

Measuring Cognitive Load:
A Comparison of Self-report and Physiological Methods

by

Stacey Joseph

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved January 2013 by the
Graduate Supervisory Committee:

Robert Atkinson, Chair
Mina Johnson
Brian Nelson
James Klein

ARIZONA STATE UNIVERSITY

May 2013

ABSTRACT

This study explored three methods to measure cognitive load in a learning environment using four logic puzzles that systematically varied in level of intrinsic cognitive load. Participants' perceived intrinsic load was simultaneously measured with a self-report measure—a traditional subjective measure—and two objective, physiological measures based on eye-tracking and EEG technology. In addition to gathering self-report, eye-tracking data, and EEG data, this study also captured data on individual difference variables and puzzle performance. Specifically, this study addressed the following research questions: 1. Are self-report ratings of cognitive load sensitive to tasks that increase in level of intrinsic load? 2. Are physiological measures sensitive to tasks that increase in level of intrinsic load? 3. To what extent do objective physiological measures and individual difference variables predict self-report ratings of intrinsic cognitive load? 4. Do the number of errors and the amount of time spent on each puzzle increase as the puzzle difficulty increases? Participants were 56 undergraduate students. Results from analyses with inferential statistics and data-mining techniques indicated features from the physiological data were sensitive to the puzzle tasks that varied in level of intrinsic load. The self-report measures performed similarly when the difference in intrinsic load of the puzzles was the most varied. Implications for these results and future directions for this line of research are discussed.

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Robert Atkinson for the inspiration to pursue my research interests in the LSRL. The experience was invaluable as he gave me the independence to explore and the encouragement and feedback necessary to grow as a researcher. I had access to a wealth of lab resources, a team of intelligent colleagues, and the opportunity for rich and meaningful discussions about measurement and data analysis. I will forever be grateful for these opportunities.

I would also like to thank my committee members, Dr. Mina Johnson-Glenberg, Dr. James Klein, and Dr. Brian Nelson for their time and support during my dissertation process. Their insight and feedback was invaluable. Additionally, I am indebted to my colleagues Dr. Mustafa Baydogan for sharing his expertise in data mining and to Dr. Lija Lin for his friendship and discussions about statistics. I also appreciate Robert Christopherson for supporting the technology in the lab.

Finally, I appreciate that this research was funded by the Office of Naval Research, Grant N00014-10-1-0143, awarded to Robert K. Atkinson.

TABLE OF CONTENTS

| | Page |
|--|------|
| LIST OF TABLES | v |
| LIST OF FIGURES..... | vi |
| CHAPTER | |
| 1 INTRODUCTION..... | 1 |
| Theoretical Framework..... | 3 |
| Overveiw of Study..... | 13 |
| 2 METHOD | 18 |
| Participants and Design | 18 |
| Problem-Solving Environment..... | 18 |
| Individual Difference Measures | 20 |
| Self-report Measure of Cognitive Load | 22 |
| Physiological Measure of Cognitive Load..... | 22 |
| Puzzle Performance Measures..... | 24 |
| Procedure | 24 |
| Scoring | 26 |
| 3 RESULTS..... | 27 |
| Self-report Ratings of Cognitive Load..... | 28 |
| Physiological Data..... | 29 |
| Puzzle Performance: Error and Time..... | 37 |

| | |
|--|------|
| CHAPTER..... | Page |
| 4 DISCUSSION | 41 |
| Research Questions..... | 41 |
| Implications..... | 46 |
| Limitations and Future Direction | 48 |
| REFERENCES | 50 |
| APPENDIX | |
| A PRACTICE PUZZLES | 76 |
| B EXPERIMENTAL PUZZLES | 78 |
| C PARTICIPANT DATA SURVEY..... | 80 |
| D COGNITIVE LOAD AND PUZZLE SELF-EFFICACY SURVEY | 83 |
| E EXIT SURVEY | 85 |
| F IRB Approval | 87 |
| G LICENSING AGREEMENT..... | 89 |

LIST OF TABLES

| Table | | Page |
|-------|--|------|
| 1. | Description of Dependent Measures used to Assess Cognitive Load | 57 |
| 2. | Puzzle Description | 58 |
| 3. | Puzzle Difficulty: Target Car and Nontarget Car Moves | 60 |
| 4. | EEG Spectral Features | 62 |
| 5. | Confusion Matrix for EEG Spectral Features | 64 |
| 6. | Emotiv Features and p Values for Predicting Puzzle | 65 |
| 7. | Confusion Matrix: Emotiv Features for Predicting Puzzle | 66 |
| 8. | EEG Spectrum Features for Predicting Difficulty Ratings | 67 |
| 9. | Confusion Matrix: EEG Spectral Features for Predicting Self-report Difficulty Ratings | 69 |
| 10. | Emotiv Features and p Values for Predicting Self-report Ratings of CL | 70 |
| 11. | Confusion Matrix: Emotiv Features for Predicting Self-report Rating of CL | 71 |
| 12. | Puzzle Errors: Means and Standard Deviations for Original and Transformed Data | 72 |
| 13. | Time to Solve Puzzle: Means and Standard Deviations for Original and Transformed Data | 73 |
| 14. | Median Values for Puzzle Errors and Time | 74 |
| 15. | Self-report Ratings of Cognitive Load | 75 |

LIST OF FIGURES

| Figure | Page |
|---|------|
| 1. Image of Practice Puzzle C with Labeled Components | 59 |
| 2. Data Mining Process | 61 |

Chapter 1

INTRODUCTION

It is common practice in cognitive load research to measure cognitive load with self-report instruments (Paas, Tuovinen, Tabbers, & van Gerven, 2003). Typically, these measures produce survey data that are indicative of participants' perception of cognitive load. There are a number of advantages of relying on self-report instruments including that they are relatively unobtrusive, easy to administer, and sensitive to changes in cognitive load (Paas, 1992). Nevertheless, there is current debate in the educational research literature concerning the construct of cognitive load as well as the reliability and validity of cognitive load measurement techniques (de Jong, 2010; Schnotz & Kirschner, 2007; Sweller, 2010; Whelan, 2007). This is due, in part, to the subjective nature of self-reports. Researchers are also exploring more objective measures of cognitive load using physiological techniques such as electroencephalography (EEG) and pupillometry (Anderson, Potter, Matzen, Shepherd, Preston, & Silva, 2011; Klingner, 2010). However, there remains a shortage of research comparing measurement methods to assess cognitive load including both traditional self-report instruments and recent advances in physiological measurement technologies. There are even fewer studies that examine multiple measurement methods in a controlled learning environment that allows for the systematic manipulation of intrinsic cognitive load.

This study was designed to simultaneously explore physiological and self-report measures of cognitive load in a controlled learning environment where one source of cognitive load was manipulated. The goal of the exploration was to determine the extent to which these measures were sensitive to fluctuations in intrinsic cognitive load and to

establish the utility of physiological sensor data to assess cognitive load. In addition, the effects of individual differences on self-report rating of cognitive load were explored and puzzle performance data (number of puzzle errors and the time spent solving puzzles) were analyzed as a means to determine how these data trended as puzzle difficulty increased. Specifically, the following research questions were addressed:

1. Are self-report ratings of cognitive load sensitive to tasks that increase in level of intrinsic load?
2. Are physiological measures sensitive to tasks that increase in level of intrinsic load?
3. To what extent do objective physiological measures and individual difference variables predict self-report ratings of intrinsic cognitive load?
4. Do the number of errors and the amount of time spent on each puzzle increase as the puzzle difficulty increases?

Discovering the relationships between physiological sensor data, self-report ratings of cognitive load, and task difficulty has potential to provide substantive validity evidence for the construct of cognitive load and to support the utility of physiological measures that produce digital signals to detect and monitor cognitive load. Using digital signals and tested algorithms to detect cognitive load in computer learning environments can function to trigger adaptations in learning content that will more effectively meet the needs of the learner. For example, when a learning system detects a state of high cognitive load it could trigger the display of a hint or provide a redirect to foundational learning content; alternatively, when a low, suboptimal, state of cognitive load is detected it could activate a feedback message of encouragement or elevate the learner to a more challenging path.

Theoretical Framework

Over the past two decades, cognitive load theory has provided a framework for research on cognitive processes, which has led to the development of instructional design guidelines (Paas, Renkl, & Sweller, 2003) and effective learning environments (Ayres & van Gog, 2009). The theory is based partially on a multicomponent working-memory model (Baddeley, 2007) that presupposes that humans have a limited working memory (WM) capacity as well as two partially independent subsystems for temporarily storing and processing different types of sensory input: the visuospatial sketchpad (VSSP) and the phonological loop (PL) (Paas, Tuovinen, Tabbers, & van Gerven, 2003). The VSSP is a storage system that integrates visual and spatial information, while the PL processes auditory information as well as visual representations of verbal information (Baddeley, 2007; Gyselinck, Jamet, & DuBois, 2008). Cognitive load theory asserts that these WM subsystems are controlled by long-term memory schema that “can act as a central executive” (van Merriënboer & Sweller, 2005, p. 149) and coordinate cognitive tasks by directing attention to relevant information (Sweller, 2005). In theory, for learning to occur, information in WM must be connected to prior-knowledge schema and converted to long-term memory storage (Sweller, 2005). Well-designed instruction and effective learning activities can facilitate this process and even expedite the construction of a learner’s incomplete long-term memory schema. However, given that WM has limited capacity and that incoming information is processed in separate temporary storage systems, a learner can experience cognitive load that may interfere with learning. Cognitive load is experienced to the extent that the WM-processing demands for a task exceed the learner’s cognitive capacity (Mayer & Moreno, 2003). It is also moderated by

an interaction between individual differences (e.g., age, expertise level, and spatial ability) and characteristics of the learning task such as format, complexity, use of multimedia, time pressure, and instructional pace (Paas, Tuovinen et al., 2003; Wouters, Paas, & van Merriënboer, 2008). For example, learners who have high prior knowledge may be able to effectively process information under conditions of increased levels of cognitive load because their cognitive schemata in long-term memory compensate or substitute for necessary WM capacity.

Sources of Cognitive Load

Cognitive load theory accounts for three different sources of cognitive load: intrinsic, extraneous and germane. Intrinsic load is caused by the inherent difficulty of the learning task and is typically determined by the number of interacting elements (element interactivity) necessary to process the task (Sweller, 2005). Element interactivity is the extent to which relevant chunks of instructional information interact in WM (Paas, Renkl et al., 2003). As the number of interacting elements increases, the intrinsic cognitive load of the task increases. For example, highly complex information that needs to be processed simultaneously for understanding would cause high intrinsic load. Extraneous load is caused by the suboptimal design of instruction and is considered detrimental to the learning process. Extraneous load can occur when multiple sources of redundant information are presented or irrelevant details are included. Sweller (2010) indicates that “element interactivity is the major source of WM load underlying extraneous [load] as well” (p. 125). Consequently, element interactivity can be attributed as the cause of both intrinsic and extraneous cognitive load. If the element interactivity is high as a result of the instructional design, it is considered to add to extraneous load; if the element

interactivity is high because of the nature or difficulty of the cognitive task, then the load is considered intrinsic. In contrast, germane load is the WM load that is essential for learning: the effortful and necessary processing that promotes development of cognitive schema (Sweller et al., 1998) while overcoming the intrinsic load inherent to the instructional information (Sweller, 2010). Learners must expend effort on the learning process to make connections with prior knowledge and alter and expand their schema in long-term memory. Because these sources of cognitive load are assumed to be additive (Sweller, 2005), the challenge for researchers and instructional designers is determining how best to create learning environments to manage element interactivity, whether it is caused by intrinsic or extraneous cognitive load, so that a learner's WM and cognitive capacity does not become overloaded.

A first step in creating effective adaptive learning environments is determining how to measure cognitive load. For instance, identifying instructional elements in a computer-based learning environment that cause high levels of cognitive load could prompt the interface to display hints or worked examples to reduce cognitive load (Ayres, 2006) and free up WM resources. However, measuring cognitive load is a complex task; cognitive load researchers have not yet agreed upon ideal measures of cognitive load, nor have they agreed upon which techniques measure which sources of cognitive load (Paas, Tuovinen et al., 2003). In the educational research literature, various subjective and objective methods have been used to measure total cognitive load (see Paas, Tuovinen et al., 2003, for a review). Recently, researchers have attempted to measure different sources of cognitive load with subjective rating scales (Ayres, 2006; Cierniak, Scheiter, & Gerjets, 2009; DeLeeuw & Mayer, 2008; Gerjets, Scheiter, Opfermann, Hesse, &

Eysink, 2009). However, this recent line of research has spurred debate. Some researchers “are in doubt whether learners will really be able to clearly distinguish different kinds of cognitive load by introspection” (Schnotz & Kürschner, 2007, p. 500) and argue that there is a need to explore “alternative approaches to the measurement of cognitive load” in order to “accurately and comprehensively measure cognitive load in the instructional design process” (Whelan, 2007, p. 4).

Measuring Cognitive Load

As research on applications of cognitive load theory in instructional design practices has evolved, so have the techniques used to measure cognitive load. In the late 1980s when researchers were beginning to study cognitive load, measurement efforts primarily entailed the use of performance-based measures such as learning time, error rate, and achievement scores (Paas, van Merriënboer, & Adam, 1994). This focus was problematic because performance measures do not take into account variability in performance due to individual differences (e.g., level of prior knowledge or aptitude) that may influence one’s experience of cognitive load. For example, as discussed above, a student with high prior knowledge may perform well on a test despite high extraneous load because prior knowledge compensates for the load and frees up WM resources.

Other well-utilized measurement techniques include both subjective and objective methods. Subjective methods primarily include self-report ratings of cognitive load, whereas objective methods include performance data from dual-task methods (e.g., employing a primary task to teach content and occupy attention while assessing performance on a secondary task) and physiological data such as eye-movement, pupil dilation, and brain-wave activity. Although a number of studies have examined self-

report measures in conjunction with dual-task methods (e.g., Brunken, Steinbacher, Plass, & Leutner, 2002; DeLeeuw & Mayer, 2008), few have explored self-report and physiological measures simultaneously. These two types of measurement techniques are described in the following sections.

Subjective techniques. Self-report methods primarily include gathering data directly from learners who rate their experience of cognitive load on a Likert-type scale. Such methods are widely used to measure cognitive load as they are relatively easy to administer, especially in authentic learning environments, and have been shown to provide an indication of cognitive workload (Ayres, 2006). Researchers typically employ a one-item assessment question where participants rate the amount of mental effort required to solve a problem on a 7- or 9-point scale. Paas, Tuovinen et al. (2003) reviewed 27 cognitive load studies dated 1992 through 2003 to determine the measurement techniques that were being used in the field. Of the studies they reviewed, 24 utilized a self-report measure; of these studies, 75% included a mental effort question on a 7-point or 9-point response scale. As an example, one of the initial studies (Paas, 1992) was designed to test a self-report measure of cognitive load and three different problem-solving conditions; participants self-reported the amount of mental effort they used to solve statistics problem on a 9-point scale from “very, very low mental effort (1) to very, very high mental effort (9)” (p. 430). Participants were asked to provide a rating of mental effort after each problem they solved during the instructional phase, as well as on the posttest.

More recent studies have used a variation of this item and rating scale. For example, Ayres (2006) utilized a question addressing the difficulty level of calculations

on a 7-point scale ranging from 1 (*extremely easy*) to 7 (*extremely difficult*). Participants were asked to provide a rating for how easy or difficult they found each calculation. Ayres indicated that “by asking students to rate task difficulty [for each calculation], students were providing an overall measure of cognitive load each time” (p. 393). DeLeeuw and Mayer (2008) conducted two studies in which they assessed cognitive load using a question about effort during the learning module and a question about difficulty at the end of the learning module. It is of note that the researchers reported a significant correlation between participants’ difficulty rating and effort rating ($r = .33, p < .01$).

Measuring cognitive load with a one-question survey item is a common practice in cognitive load research; however, the difference between assessing cognitive load with a question about *difficulty* versus a question about *mental effort* is unclear. There is evidence from DeLeeuw and Mayer (2008) that data gathered using both items are significantly correlated and that perhaps the items can be used interchangeably. As noted earlier, self-report methods require participants to be introspective about their cognitive processes or have metacognitive insight into the effort required to complete learning tasks. Rating accuracy may be affected by participants’ ability to be introspective. More reliable and valid results may be attained by asking participants to judge the difficulty level of a problem as opposed to introspecting or making judgments about their cognitive processes. It also may be useful to use additional items as indicators of cognitive load, such as levels of stress and frustration. Recently, researchers have explored using a multiquestion assessment tool (Cierniak et al., 2009; Gerjets et al., 2009).

Objective techniques. Even though subjective measures of cognitive load are widely used and are generally unobtrusive, physiological methods have also been used to

measure cognitive load. Eye-tracking techniques and EEG data offer less-obtrusive empirical techniques for exploring cognitive load (Amadiou, van Gog, Paas, Tricot, & Mariné, 2009). Data from the eye-tracker have been used in various disciplines to study attention and cognitive processing in the form of eye-gaze patterns (Duchowski, 2002) and pupil dilation (see Beatty, 1982, for a review). For example, cognitive overload and attention have been studied via pupil dilation in schizophrenic patients (Minassian, Granholm, Verney, & Perry, 2004). Results showed that there is increased pupil dilation in both patients and nonpatients for a complex task compared to an easy task. Pupil changes during mental activity have been related to task difficulty, as “the pupillary response appears to reflect the information processing load placed on the nervous system by cognitive tasks (Andreassi, 2000, p. 357). Other researchers have obtained similar experimental results among college students to support a link between cognitive effort and increased pupil dilation. Verney, Granholm, and Dionisio (2001) found participants’ pupillary dilation responses to be significantly more pronounced during task performance (cognitive load) compared to passive performance. For a visual search task where difficulty was manipulated by varying the number and type of distractors, Porter, Troscianko, and Gilchrist (2007) concluded that pupil dilation increased as the difficulty of the search task increased. Another study conducted by researchers in computing science used eye tracking to study task-evoked pupillary responses and cognitive load (Klingner, Kumar, & Panrahan, 2008). Exploring eye-tracking data, pupil-dilation data, and self-report data in one study may lend insight into how differentially sensitive these measures are to cognitive load, and it may provide information about the predictive validity of eye-gaze and pupil-dilation data.

Additionally, EEG is a well-used noninvasive neuroimaging technique designed to capture continuous brain-wave activity, such as alpha, beta, and theta waves. It has been established in the literature that EEG data vary predictably in response to changes in cognitive stimuli (Anderson, Potter, Matzen, Shepherd, Preston, & Silva, 2011; Gevins & Smith, 2003) and WM load (Klimesch, Schack, & Sauseng, 2005; Klimesch, Schimke, & Pfurtscheller, 1993), making EEG an appropriate choice for measuring cognitive load (Antonenko, Paas, Grabner, & van Gog, 2010). For example, research evidence suggests brain-wave activity in the alpha and theta bands is reactive to increases in task difficulty (Gevins & Smith, 2003; Gevins et al., 1998; Smith, Gevins, Brown, Karnik, & Du, 2001). Specifically, the lower-frequency alpha signals, from 8 to 10 Hz, tend to desynchronize or become lower in power as task difficulty increases, while theta signals, from 4 to 7 Hz, tend to synchronize or increase in power. EEG supplies a continuous measure of cognitive load that provides an opportunity for researchers to gather and analyze fluctuations in a stream of data over time as opposed to the few data points derived from self-report techniques.

The device used to acquire EEG data for this study was an Emotiv EPOC, a neuro-signal wireless headset. It is designed as a low-cost video game controller that contains proprietary software algorithms created to measure affective constructs such as (a) engagement, (b) instantaneous excitement, (c) long-term excitement, (d) frustration or boredom, and (e) meditation. The algorithms are not publicly available, which leaves unclear the specific relationships between the constructs and (a) increases in intrinsic load for the tasks and (b) self-report difficulty ratings of those tasks. Consequently, both the affective constructs and the raw EEG data were considered independently as components

of the physiological data set that was analyzed with data-mining techniques to uncover prediction models that predict task difficulty and self-report ratings of difficulty.

Individual Differences

Cognitive load theory is based upon a multicomponent WM model, and individual differences have the potential to moderate the capacity of WM. In order to create a controlled experiment and systematically manipulate one source of cognitive load, it is necessary to explain or account for individual difference variables such as prior knowledge, WM capacity, spatial visualization, and self-efficacy. Prior knowledge, as in the expertise reversal effect, has been found to affect task performance and perception of cognitive load (Kalyuga, 2007; Kalyuga, Chandler, & Sweller, 1998; Kalyuga, Chandler, & Sweller, 2000; Lee, Plass, & Homer, 2006; Kalyuga, Ayres, Chandler, & Sweller, 2003; for a review, see van Merriënboer & Sweller, 2005). The expertise reversal effect attempts to explain how the effectiveness of an instruction format can depend upon the extent of a learner's prior knowledge. For example, learners who have a high level of domain-specific prior knowledge may perform worse on a task compared to learners with lower-level domain knowledge (Kalyuga, 2005). This effect was well demonstrated by Kalyuga et al. (1998) in a split-attention experiment: Novice students who studied integrated diagrams (text within diagram) learned more compared to novice learners who studied the same information separately. The advantage of the integrated instructional format was neutralized as novice students became more expert, and then it was eventually reversed to become a detriment. In fact, students with higher domain knowledge learned more with a diagram only (and no text) because the text became redundant information (van Merriënboer & Sweller, 2005) that interfered with learning.

Similarly, spatial visualization (Höffler, 2006; Mayer & Moreno, 2003; see Höffler, 2010, for a meta-analysis) and, to a lesser extent, WM capacity (van Gerven, Paas, van Merriënboer, & Schmidt, 2002) have been investigated to explain learner differences on task performance. As noted by Ekstrom, French, and Harman (1976), spatial visualization is “the ability to manipulate or transform the image of spatial patterns into other arrangements” (p. 173) via serial operations in short term visual memory (Carroll, 1974). Common measures of spatial-visualization ability include tasks such as paper-folding and mental rotation. Höffler (2010) conducted a meta-analysis of 27 experiments published between 1994 and 2009 to determine the effects on learning outcomes when high-spatial-ability learners and low-spatial-ability learners studied various types of visualizations. Höffler found an overall mean effect size of $r = 0.34$, which indicated a medium effect size for high-spatial-ability learners. Specifically, learners with higher spatial ability who studied visualizations had better learning outcomes than learners with lower spatial ability.

Despite the fact that cognitive load theory is based on a WM model (Baddeley, 2007) and the experience of cognitive load is contingent upon available WM capacity, with few exceptions (e.g., van Gerven et al., 2002) WM capacity is not often included as a covariate in cognitive load research (de Jong, 2010). WM capacity has been shown to be related to performance in reasoning and reading comprehension (Engle, Cantor, & Carullo, 1992); learning in mobile (Doolittle & Mariano, 2008), hypertext (DeStefano & LeFevre, 2007), and e-learning (Tsianos, Germanakos, Lekkas, Mourlas, & Samaras, 2010) environments; and self-regulation of emotion (Schmeichel, Volokhov, & Demaree, 2008). Widely used measures of WM capacity include span tasks, such as operation-span

and reading-span tasks (Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle, 2005). Span tasks utilize a dual-task methodology to measure the ability to control attention and thought by forcing WM storage while distractors are processed simultaneously (Conway et al., 2005).

Self-efficacy is also a relatively unstudied construct within the cognitive load research literature, and it is included as an exploratory measure. Perceived self-efficacy is a judgment about one's capability (Bandura, 1977) to accomplish a task. As Pajares (2002) noted, "people's accomplishments are generally better predicted by their self-efficacy beliefs than by their previous attainments, knowledge, or skills. Of course, no amount of confidence or self-appreciation can produce success when requisite skills and knowledge are absent" (p. 16). It follows that self-efficacy may play a role in predicting performance. For example, Pajares and Miller (1994) conducted an experiment with undergraduate students ($N = 350$) to test the role of self-efficacy in solving mathematics problems. Their results showed that math self-efficacy was a better predictor of problem solving than prior knowledge in math. Likewise, Narciss (2004) conducted two experiments, and results revealed that motivation and achievement depended upon both self-efficacy and type of feedback.

Overview of Study

The overarching purpose of this study was to attempt to establish the utility of measuring cognitive load with physiological measures by assessing the responsiveness of self-report and physiological measures in an experimental environment where intrinsic cognitive load (element interactivity) was systematically manipulated. Secondly, the effect of individual difference variables on self-report ratings of cognitive load was

explored and puzzle performance data were analyzed. As mentioned previously, there were four research questions:

1. Are self-report ratings of cognitive load sensitive to tasks that increase in level of intrinsic load?
2. Are physiological measures sensitive to tasks that increase in level of intrinsic load?
3. To what extent do objective physiological measures and individual difference variables predict self-report ratings of intrinsic cognitive load?
4. Do the number of errors and the amount of time spent on each puzzle increase as the puzzle difficulty increases?

In order to answer these questions, an experimental environment was created to isolate and systematically manipulate levels of intrinsic load. The effect of the manipulation was assessed simultaneously using different measurement techniques: a self-report measure and physiological measures that utilized eye-tracking and EEG technology . In order to isolate intrinsic load, a computer-based logic puzzle environment was used, with puzzles that required very little instruction. This strategy was similar to Ayres (2006), in that there was a deliberate attempt to exclude instructional material in order to minimize effects of extraneous load and to use puzzle tasks that varied only in surface features. Level of intrinsic load was manipulated by varying the number of interacting elements presented across four puzzles. In this case, intrinsic load or element interactivity was defined as the interactions among the puzzle pieces as they were manipulated toward the goal state (see “Puzzle difficulty” below, in the method section, for a detailed description). Three types of process measures were used to explore the

impact of increasing the level of intrinsic load across puzzle tasks: self-report, physiological, and puzzle performance. . Self-report measures were subjective cognitive-load ratings measured after each puzzle. Physiological measures included eye-tracking (fixation durations and pupil data) and EEG (brain activity and Emotiv EPOC construct data). Puzzle performance measures included (a) time to complete each puzzle task and (b) number of moves required to complete each puzzle task over and above the required minimum (errors). In an attempt to control for individual differences, WM capacity, spatial visualization ability, puzzle prior knowledge, and puzzle self-efficacy were measured. A description of the variables is summarized in Table 1.

Pilot Study

Prior to examining these research questions, we conducted a pilot study to validate that perceived cognitive-load ratings were sensitive to manipulations of element interactivity (i.e., intrinsic load) in the logic puzzle environment (Schink Joseph & Atkinson, 2010). The pilot addressed the question of whether self-report ratings of cognitive load are sensitive to tasks that vary in levels of intrinsic load. The study revealed that participant ratings of perceived cognitive load increased as the element interactivity of the puzzles increased; the effect was significant when the puzzle with the least element interactivity (the easiest puzzle) was compared to each of the other puzzles that increased in element interactivity. This finding indicated that participant ratings of cognitive load were sensitive to manipulations of element interactivity (i.e., intrinsic load) within the puzzle tasks.

Interestingly, pairwise comparisons of difficulty ratings on the three more-difficult puzzles (requiring more moves to complete) were not significant. It appears that

in this puzzle environment, subjective ratings are significantly sensitive to manipulations of element interactivity under specific conditions. When differences in element interactivity are most variable—for example, the difference between an easy puzzle compared to more difficult ones—these differences can be detected by self-report ratings. However, perhaps when element interactivity reaches a certain threshold, as in the three more difficult puzzles, self-report ratings of cognitive load are less sensitive.

Data Mining

In this experiment, analyzing physiological sensor data with conventional inferential statistics is problematic, as it includes multiple heterogeneous variables and very large data sets. Consequently, an alternate data-analysis method is necessary to efficiently discover the most important variables for predicting the difficulty level of the puzzle (Research Question 2) and for predicting self-report difficulty ratings for each puzzle task (Research Question 3), while avoiding issues of multicollinearity and theoretical assumptions required for inferential statistics. Data mining is an appropriate technique to accomplish this result; it is a procedure that utilizes machine-learning algorithms to identify useful patterns of information in large, complex data sets and to make predictions about specific attributes (Tan, Steinbach, & Kumar, 2006). Such patterns, or classifications of data, can be extracted via data-mining techniques from even “large, noisy, messy data sets” (Nisbet, Elder, & Miner, 2009, p. 17).

There are a number of examples in educational research where data-mining classifiers (e.g., linear regression, decision trees, Bayesian classifiers, and neural networks) have been used to predict outcome variables such as academic success, course outcomes, metacognitive skills, and other factors that impact learning (e.g., motivation,

engagement; see Hamalainen & Vinni, 2011, for examples). The classifier Random Forest (RF; Breiman, 2001) is an ensemble of decision trees where “each tree is constructed using a different bootstrap sample from the original data” (Baydogan, Runger, & Tuv, 2011, p. 6). Breiman (2001) noted strengths of this particular classifier: (a) It gives useful internal estimates of error, strength, correlation, and variable importance; (b) overfitting is not an issue; and (c) it accommodates data sets with many variables where each of the variables contains little information. RF can also simultaneously handle both categorical and continuous types of data. These strengths are important to utilize when aggregating and analyzing the thousands of data points typically extracted during eye tracking, pupillometry, and EEG studies.

Participants and Design

Participants were 56 undergraduate college students (37 women and 19 men) from a large southwestern university who ranged in age from 18 to 36 ($M = 21.5$, $Mdn = 20$). Other self-identified demographic data from the sample indicated that their primary spoken language was English (98%, 2% Spanish). The ethnic composition of the sample was 57% White, 18% Hispanic, 7% African American, 7% Asian/Pacific Island descent, and 10% indicating they belonged to another, or more than one, racial or ethnic group. Students were recruited from pools of undergraduates who attended either an introductory psychology class or an introductory educational technology class. They were paid \$20 to participate.

To examine the research questions outlined above, a one-way within-subjects design was used. The factor was puzzle difficulty, with four levels of difficulty, or level of intrinsic cognitive load. Puzzle order was counterbalanced according to a Latin square design to ensure that each puzzle appeared in each position one time and was never preceded or followed by the same puzzle. This design was chosen to minimize the influence of puzzle order on puzzle performance. Participants were randomly assigned to one of the four counterbalanced conditions.

Problem-Solving Environment

The computer environment consisted of puzzle tasks and electronic measures that were automated by a program built in QuickKeys. The puzzle tasks, developed by Hearn (2009), consisted of seven modified Subway Shuffle puzzles designed for Mac OS X—

three practice puzzles and four experimental puzzles. Subway Shuffle, a one-player game, is a sequential-movement puzzle that can be considered a variant of a sliding-block puzzle (Hearn, 2006). A sliding-block puzzle essentially consists of puzzle pieces that are any shape, are contained within a defined space, and move independently from each another in a sliding fashion from one position to another; the purpose of the puzzle is to arrange the pieces into a predefined pattern or to move a certain piece to a specified position (Hordern, 1986).

Puzzle description. Each Subway Shuffle puzzle consisted of movable colored tokens (subway cars), stations, and different-colored stationary subway tracks (see Appendices A and B for an image of each puzzle). The goal of the puzzles was to slide the red subway car (target car) to unoccupied stations along the red track to reach the final destination marked with a red ring (see Figure 1 for labeled example). Each puzzle contained only one target car, four or five cars of other colors, six or seven stations, and six or nine segments of colored track. There were three or four colors (for cars and tracks) used in each puzzle: blue, red, yellow, and green (see Table 2, for summary). In order to advance the target car, it was necessary to move other subway cars that occupied stations blocking the target car's path. Only one subway car could move at a time, and a car could move only on the track that was the same color as the car (i.e., a blue subway car could move only along a blue color track) to an unoccupied station.

Puzzle difficulty. Puzzle difficulty increased as each of the following elements were altered: number of moves required to complete the puzzle and quantity of subway cars, tracks, colors, and subway stations. Varying these elements caused an increase in the number of moves required of nontarget subway cars relative to moves required of the

target car and thereby increased the difficulty of the puzzle. For example, for each of the puzzles, the target car was required to move 3–5 times to reach its destination in the fewest number of moves, while the nontarget subway cars were required to move 8–18 times (see Table 3). The experimental puzzles (listed in order from easier to harder) required the following number of nontarget car moves: Puzzle 1 – 8, Puzzle 2 – 10, Puzzle 3 – 14, and Puzzle 4 – 18. An increase in the number of required nontarget car moves (element interactivity) caused an increase in working memory load as puzzle solvers must create a spatial mental path that leads to the goal state and must simultaneously select correct moves and inhibit moves that incorrectly appear to advance the target car to its destination. The solution path is determined by the number of required nontarget cars moves, and it becomes more difficult to solve as the number of required nontarget car moves increases. Performance on this type of puzzle is similar to what is required to solve the Tower of Hanoi task (Handley, Capon, Copp, & Harper, 2002).

Individual Difference Measures

Working memory capacity. WM capacity was measured using an operation span task initially designed by Turner and Engle (1989). The task was adapted by Lewandowsky, Oberauer, Yang, and Ecker (2010) for electronic administration via the Psychophysics Toolbox (Brainard, 1977; Pelli, 1997) in Matlab. A structural equation model analysis indicated that the factor loading for the electronic version of the OS task and WM is .77 (standardized estimate). For a series of 15 trials:

- Participants viewed a sequence of solved mathematics equations followed by consonants (excluding Y and Q) and then judged the accuracy of each equation and remembered the consonants for later recall.

- After viewing a trial, participants indicated if the equation was solved correctly (yes or no) and then recalled the consonants, one at a time, in the order they were displayed.
- The length of the consonant list varied from four to eight (consonants lists were not repeated) and there were three trials per list length, with a total of 15 trials.
- Trial order, consonants, and equations were randomized into one order and presented to all participants.
- There was no time limit for recall, 500 ms between trials, and a self-paced break after every three trials (Lewandowsky et al., 2010, p. 573).

Spatial visualization. The paper-folding test (Ekstrom, French, & Harman, 1976) was used to assess spatial visualization ability. The test included a direction sheet and two paper-folding activity sheets that contained 10 items each. Each item was organized into two corresponding columns, a left-hand (LH) column and a right-hand (RH) column. For each item, the LH column contained images of a piece of square paper in a sequence of two to four folds; the final folded image was hole-punched with one hole. The RH column contained five images of the paper as it might appear unfolded. Participants circled the image that correctly identified the piece of paper from the LH as it would appear unfolded. There was a 3-minute time limit for each activity sheet.

Prior knowledge. The Participant Data Survey (Appendix C) was used to gather participant data regarding demographics, vision issues, and puzzle prior knowledge. Prior-knowledge questions included items designed to gather information about the types of puzzles participants played and the frequency with which they played. There was also an image of a puzzle similar to the experimental puzzles and a corresponding question

used to determine if participants had previous experience playing the experimental puzzles.

Self-efficacy. The Cognitive Load and Puzzle Self-Efficacy Survey (Appendix D) captured puzzle self-efficacy data. There were two questions, modeled after Bandura (2006), for which participants rated their response on a 9-point scale: (a) “How confident are you that you can solve more puzzles like this?” and (b) “How certain are you that you can solve more puzzles like this?”

Exit survey. The Exit Survey (see Appendix E) was used to obtain opinion data about the puzzle environment. There were five questions designed to explore participants’ perception of overall task difficulty, and participants rated their responses to each on a Likert-type scale. These data are supplemental and will be used for exploratory analyses.

Self-report Measure of Cognitive Load

The Cognitive Load and Puzzle Self-Efficacy Survey (see Appendix D) contained three cognitive load questions, and participants rated their response to each on a 9-point scale. The measure of cognitive load was similar to Ayres (2006): “How difficult was this puzzle to solve?”

Physiological Measures of Cognitive Load

Eye tracking. A 24-inch, 60Hz Tobii T60 XL Eye Tracker monitor was used along with Tobii Studio software (Tobii Technologies, n.d.) to record eye tracking data. The computer problem-solving environment was sent from a 13-inch Apple MacBook Pro to the eye tracker. The eye-tracking software recorded pupil dilation data at rate of 16

samples per second for the left and right pupil as well as the number, duration, and locations of eye fixations.

EEG. The Emotiv EPOC wireless headset was used to collect EEG data continuously from 14 scalp locations (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4) at a rate of 128 samples per second (see Pivik, Broughton, Coppola, Davidson, Fox, & Nuwer, 1993, for location description). In addition to recording a participant's head movement data (GYROX, GYROY), the EEG brain wave activity for each channel was sent via a logger software component (Gonzalez-Sanchez, Chavez-Echeagaray, Atkinson, Burleson, 2011) to two separate log files. One log file stored a record of raw brain wave data for each channel as microvolts, and one log file stored the corresponding affective data, transformed by algorithms in the integrated TestBench software program, as the following constructs: excitement, engagement, boredom, frustration, and meditation. The constructs are described below (Emotiv, 2010):

- Engagement: experienced as alertness and the conscious direction of attention towards task-relevant stimuli. It is characterized by increased physiological arousal and beta waves along with attenuated alpha waves. The opposite pole of this detection is a sign of lack of engagement.
- Instantaneous excitement: experienced as an awareness or feeling of physiological arousal with a positive value. Excitement is characterized by activation in the sympathetic nervous system, which results in a range of physiological responses including pupil dilation, eye widening, sweat gland stimulation, heart rate and muscle tension increases, and blood diversion.

- Long-term excitement: experienced and defined in the same way as instantaneous excitement, but the detection is designed and tuned to be more accurate when measuring changes in excitement over longer time periods, typically measured in minutes.
- Frustration: experienced as a disconnection from what is expected and what is actually happening. Also, frustration is experienced when the difficulty level of a task is increased disproportionate to the skill level of a participant.
- Meditation: experienced as calm and a clearness of the mind. It is considered similar to sleep, but in a conscious state.

Puzzle Performance Measures

Time on puzzle. The eye tracking software recorded the time participants took to complete each puzzle.

Puzzle errors. The number of moves it took participants to solve each puzzle was recorded manually. The number of errors for each puzzle was calculated by subtracting the minimum number of moves required to solve the puzzle from the total number of moves it took for participants to complete the puzzle. For example, 15 moves were required to solve Puzzle 2; the number of errors recorded for a participant who took 22 moves to solve this puzzle was 7.

Procedure

Study participants completed the 1-hour experiment in a computer lab on the eye-tracking computer monitor. The automated puzzle environment was run on an Apple MacBook Pro connected through a switching hub to a PC. The researcher administered the study and was present for the duration of the experiment. Each participant was

randomly assigned to one of the experimental conditions according to a master list of ID numbers that preserved the anonymity of the participants and identified the order of the experimental puzzles.

Each participant was asked to sign a consent form. After the form was signed, participants were given information about the study. They were told, for example, that they would complete a series of puzzles and surveys on the eye-tracking monitor while wearing an EEG headset.

After a short explanation of the study, each participant completed the spatial ability test. Next, they were fitted with the EEG headset as the researcher explained the physical attributes of the headset as well as the form of the data collected. To ensure optimum signal transmission, each sensor was adjusted until the headset visualization tool for each sensor showed a green light. When the headset was properly fit, the participant was seated at the eye-tracking monitor and his/her eye movements were calibrated using Tobii calibration software. When calibration was satisfactory, the researcher started the automated program that corresponded to the randomly assigned ID number and began a recording for EEG data log files and the eye-tracking data. The sequence of activities was launched in the following order: (a) the operation span task, (b) Participant Data Survey, (c) Practice Puzzles 1, 2, and 3, (d) Cognitive Load and Puzzle Self-Efficacy Survey, (e) Problem-Solving Puzzles 1, 2, 3, and 4, presented in the randomly assigned order, each alternating with the Cognitive Load and Puzzle Self-Efficacy Survey, then finally (f) the Exit Survey. When the participant completed the final activity, the experimenter terminated the EEG and eye-tracking recordings, provided a debriefing explanation of the experiment, and answered the participant's questions.

Scoring

Working memory capacity. The OS task used to measure WM was scored in accordance with Lewandowsky et al. (2010). Partial credit was awarded for each of the consonants remembered correctly for each of the 15 trials. For example, for a list length of three, a score for two correctly remembered consonants was $2/3$. The total OS score for each participant was calculated as an average across the 15 trials. The minimum possible score was 0, and the maximum possible score was 1.

Paper-folding test. The paper-folding test, used to measure spatial-visualization ability, was scored with a correction-for-guessing formula: number of correct responses minus number of incorrect responses times $1/(n - 1)$ where n is the number of response options (Ekstrom et al., 1976). Because there were five response options, 1 point was given for each correct response, and .25 was subtracted for each incorrect response. The minimum possible score was 0, and the maximum score possible was 20.

Chapter 3

RESULTS

The results of analyses for, self-report, physiological, and puzzle performance data were reported for each of the following research questions:

1. Are self-report ratings of cognitive load sensitive to tasks that increase in level of intrinsic load?
2. Are physiological measures sensitive to tasks that increase in level of intrinsic load?
3. To what extent do objective physiological measures and individual difference variables predict self-report ratings of intrinsic cognitive load?
4. Do the number of errors and the amount of time spent on each puzzle increase as the puzzle difficulty increases?

The puzzle performance and self-report data for each of the four puzzles were inspected with SPSS functions for accuracy, missing values, and multivariate assumptions. These data included (a) number of errors, (b) amount of time taken to complete each puzzle, and (c) self-report ratings of cognitive load (CL). This inspection revealed that time and error data were substantially skewed and kurtotic. As a result, these data were logarithmically transformed. Mahalanobis distance was used to evaluate multivariate outliers with $p < .001$; no statistically significant outliers were found. A subsequent evaluation of multicollinearity conducted using SPSS indicated multicollinearity was not evident. Additionally, paper-folding scores, the measure of spatial visualization ability, were analyzed for gender differences and none were found.

Self-report Ratings of CL

Self-report ratings of CL were analyzed to answer the first research question: Are self-report ratings of cognitive load sensitive to tasks that increase in level of intrinsic load? These analyses were also an attempt to replicate the pilot study results (Schink Joseph & Atkinson, 2010) reported previously and to provide additional evidence that perceived CL ratings were sensitive to manipulations of element interactivity (i.e., intrinsic load) in the logic puzzle environment. A one-way within-subject ANOVA was conducted to evaluate the effect of varying the number of interacting elements across four puzzles on self-reported ratings of CL. The means and standard deviations for the self-report difficulty rating are reported in Table 15. The results of the test were statistically significant, Wilks' lambda = .34, $F(3,51) = 33.28$, $p < .001$, multivariate $\eta^2 = .66$.

Follow-up analyses were conducted for the six-pairwise comparisons among the CL ratings on the four puzzles to determine which of the average puzzle difficulty ratings were statistically significant. A Holm's sequential Bonferroni procedure was used to control for Type I error with alpha set at .008 (.05/6). The results indicated that as the number of interacting elements in the puzzles increased, participants progressively endorsed higher ratings of CL. This effect was statistically significant when the CL rating of Puzzle 1, the puzzle with the least element interactivity—the easiest puzzle—was compared to the CL ratings of each of the other puzzles that had more element interactivity, Puzzle 2, 3 and 4, $p < .01$. The comparison was also significant when Puzzle 2, an easier puzzle, was compared to Puzzle 4, the most difficult puzzle, $p < .01$. While these finding indicated participant ratings of CL were sensitive to variations in element interactivity (i.e., intrinsic load), they also suggest that CL ratings were not particularly

sensitive to differentiating puzzles that were somewhat similar in difficulty. For example, the differences between the mean CL rating for Puzzles 2 and 3 and for Puzzles 3 and 4 were not statistically significant.

Physiological Data

EEG wave data, Emotiv construct data, and pupil data were analyzed with data-mining techniques to answer the second research and third research questions: Are physiological measures sensitive to tasks that increase in level of intrinsic load? To what extent do objective physiological measures and individual difference variables predict self-report ratings of intrinsic cognitive load? As mentioned previously, the goal of data mining was to discover the variables, or features, that were most important for predicting the difficulty level of the puzzle tasks and for predicting the self-report ratings of CL.

Prior to analysis, physiological data were initially screened, reviewed and prepared for data mining. These data included the EEG wave and construct data from the Emotiv headset and the fixation and pupil data from the eye tracker. This process led to the elimination of 12 participants due to lost data files from technical difficulties with the eye-tracking and Emotiv software, leaving 44 cases for analysis. As noted previously, the physiological data streams from the Emotiv headset and the eye-tracking software were stored in separate log files and sampled at different rates of time. For each participant, there were two Emotiv log files sampled at 125 samples per second and one eye-tracking file sampled at 16 samples per second. To analyze the time-series data, it was necessary to link each timestamp in the eye-tracking file with the corresponding timestamps in the two EEG files and compile the physiological data into one file for each participant. This

was accomplished with software (Gonzalez-Sanchez, Chavez-Echeagaray, Atkinson, & Burleson, 2011) that generated 44 combined data files, one for each participant.

Generally, the procedure for using a data-mining technique to build prediction models includes the following steps: data preprocessing, feature extraction, and feature selection using decision-tree construction (see Figure 2). The procedure used to accomplish each of the steps with these data is reported in its entirety in Baydogan, Runger, and Atkinson (in press); below is a brief explanation from their research.

- Data preprocessing: Raw data were preprocessed to make the data more appropriate for data mining. Raw data typically contains noise, or artifacts, from sources other than the target. For example, raw EEG data from the Emotiv headset contained noise from facial muscle movements, eye movement, and eye blinks. Additionally, because the Emotiv headset was designed as a video game control device and not as high-quality research instrument, it is unclear each node captured more noise than is typical (Campbell, 2010). Techniques such as filtering and dimensionality reduction were used to reduce noise and separate artifacts from the data.
- Feature extraction: Distinctive features or variables in the data set were generated by transforming EEG wave data from microvolts into a wave-frequency domain using a fast Fourier transform (FFT) algorithm. To further characterize the wave-frequency data, each of the brain waves frequencies (alpha, theta, and beta) detected at each of the 14 headset nodes was represented by a mean, a minimum, and a maximum wave value. This combination created a feature set that included 126 variables. The Emotiv construct data consisted of a feature set of 35 variables.

There were five measures of variability (variance, minimum, maximum, skewness and kurtosis) and two measures of central tendency (mean and median) for each of the five constructs. Although measures of variability and central tendency do not contain information about how the data behave over time, they do provide information about the characteristics of the signal distribution. Pupil data was extracted as left and right eye data for pupil size, fixation duration, and pupil location creating a feature set of six variables.

- Feature selection and decision tree construction: Ensembles of decision trees were used to select relevant features or variables to predict the difficulty level of the puzzle and to predict the self-report difficulty ratings of each puzzle. To evaluate the performance of the prediction model, the number of cases correctly and incorrectly predicted by the model were calculated and displayed in a confusion matrix (Tan, Steinbach, & Kumar, 2006). The correctly predicted cases aggregate in the cells on the diagonal of the confusion matrix from the upper-left corner to the lower-right corner, and incorrectly predicted cases deviate from the diagonal. In general, the cases inaccurately predicted by the model appear in cells farther from the diagonal. The classification method provided a measure of variable importance that is useful for identifying a subset of features, from many, that are most relevant for the prediction tasks. However, the method does not produce a statistical criterion for each feature that can be evaluated with a significance test to determine which variables in the prediction model are statistically significant predictors. Consequently, a statistical criterion was calculated for each feature in accordance with the procedure described in Tuv, Borisov, Runger, and Torkkola

(2009). Tuv et al. used an algorithm (called ACE) to remove irrelevant features from the data set by generating a subset of random artificial features and variable importance scores to which true relevant features were compared. “Higher variable importance scores are expected from a true relevant variable than from an artificially generated contrast variable” (p. 1246). With a sufficient amount of data, importance variables can be selected from among those that have statistically significantly higher variable importance scores than the contrast variables. Paired *t*-tests evaluating the difference between values for the artificial and true relevant variables were used as a test of significance.

Predicting Puzzle Type. EEG wave data, Emotiv construct data, and pupil data were analyzed to answer the second research question: Are physiological measures sensitive to tasks that increase in level of intrinsic load? Results are reported for each of the three sets of physiological measures.

EEG data analysis. These analyses were also conducted, in part, to replicate findings in the literature indicating that alpha waves decrease and theta waves increase as a task becomes more difficult. Such results were intended to validate that the EEG data captured by the Emotiv software were useful for differentiating tasks that varied in difficulty. As noted in Baydogan, Runger, and Atkinson (in press), below is a brief description of the procedure for each of the data-mining steps with these data:

- *Preprocessing.* The raw EEG signal data were acquired from participants and initially stored in the EEG log file as units in microvolts. To remove artifacts and reduce signal noise, the signal data were low-pass filtered and transformed using independent component analysis (ICA). ICA was used to separate the original

signal from noise. Subsequently, the noise was filtered and the clean signal data were reconstructed.

- *Feature extraction.* As mentioned previously, the FFT algorithm was used to transform the EEG signals into wave-frequency data. The values for the wave data were computed by band-pass filtering the signals within specific frequency ranges for alpha (8–12 Hz), beta (13–30 Hz), and theta (4–7 Hz). Thereafter, the values were squared and an average was calculated.
- *Feature selection and decision tree construction.* A Random Forest classification algorithm was applied to the EEG spectrum information for 44 participants as a method to classify the difficulty level of the puzzles. Of the 126 features from the EEG signal data, 29 features had variable importance values and p values > 0 . Each feature is described in Table 4. The confusion matrix depicting the classifications is displayed in Table 5. These results show the percentage of the observed samples that were classified, or predicted, by the algorithm containing the 29 features. The puzzle classifications ranged in accuracy from a low of 45% for Puzzle 4 to a high of 71% for Puzzle 3. The algorithm appeared to provide better classification accuracy of Puzzles 1, 2, and 3 than Puzzle 4. For example, 55% of the samples from Puzzle 1, the puzzle with the least number of interacting elements, were classified accurately as Puzzle 1. None of the Puzzle 1 samples were classified as Puzzle 4, the puzzle with the most interacting elements, which indicates that the classification algorithm distinguished well between the easiest puzzle and the most difficult puzzle. The algorithm also performed well for accurately classifying Puzzles 2 and 3. However, for Puzzle 4, 45% of the samples

were accurately classified as Puzzle 4, and 45% of the samples were *inaccurately* classified as Puzzle 3. This suggests that when the algorithm was applied to the Puzzle 4 samples, the samples were classified equally as Puzzle 3 and Puzzle 4, indicating that the algorithm did not distinguish Puzzle 4 samples well from Puzzle 3. It is likely there were some similarities between the two puzzles that reduced the accuracy of the prediction model, a possibility that is revisited in more depth in the discussion section.

Emotiv construct data analysis. Raw data for each Emotiv affective construct were analyzed to create models with construct features that predicted tasks varying in level of intrinsic load.

- *Preprocessing.* To reduce noise in the affective construct signals generated by the Emotiv algorithms, data were smoothed by calculating an average for observed values in a 500 ms window. In other words, averages were calculated and recorded every 500 ms beginning from time 0. These average values were used for further analysis.
- *Feature extraction.* As mentioned previously, a feature set of 35 variables was used to describe the construct data. It consisted of five measures of variability (variance, minimum, maximum, and skewness and kurtosis) and two measures of central tendency (mean and median) for each of the five constructs.
- *Feature selection and decision-tree construction.* A Random Forest classification algorithm was applied to the Emotiv construct data for 44 participants to classify the difficulty level of the puzzles. In total, 14 of the 126 features extracted from the Emotiv construct data had p values > 0 . Each feature is described in Table 6. Of

the four puzzles, the algorithm was most accurate predicting puzzle samples from Puzzles 1, 2, and 4 (see Table 7 for classification accuracy). Of the Puzzle 3 samples, 23% were classified as Puzzle 3, and 36% were classified as Puzzle 2 and 36% as Puzzle 4. These findings indicate that the algorithm for the Emotiv construct data did not distinguish Puzzle 3 samples well from Puzzles 2 and 4.

Pupil data analysis. An analysis was conducted to determine if any features from the pupil-dilation data predicted the difficulty level of the puzzle. Unlike the data analysis of the raw EEG and Emotiv constructs, no features were extracted.

Predicting self-report ratings. EEG wave data, Emotiv construct data, and pupil data were analyzed to answer the third research question: To what extent do objective physiological measures and individual difference variables (spatial ability, WM capacity, and self-efficacy) predict self-report ratings of intrinsic cognitive load? The relationship between the average values for the two self-report measures, self-efficacy rating and difficulty rating, was explored for each of the puzzles to determine the usefulness of including the self-efficacy rating as a predictor in the model. Moderate to high correlations between the predictor variable (self-efficacy rating) and the outcome variable (difficulty rating) was found across the puzzles. Consequently, self-efficacy rating was eliminated from the model.

EEG data analysis. Below is a brief description of each step used in the data-mining process for these data (Baydogan, Runger, & Atkinson, in press):

- *Preprocessing and feature extraction.* The raw EEG signal data were preprocessed and the features were extracted as previously reported.

- *Feature selection and decision-tree construction.* A Random Forest classification algorithm was applied to the (a) EEG spectrum information, (b) paper-folding scores, and (c) WM capacity scores for 44 participants to classify the self-report rating of difficulty. The self-report ratings of difficulty were aggregated across puzzles generating a total of 176 responses. From the EEG signal data, 37 features were extracted that had variable importance values and p values > 0 . None of the features from the distributions of paper-folding scores or WM capacity scores were identified as important. Each EEG feature is described in Table 8. The confusion matrix depicting the classifications is displayed in Table 9. These results show the percentage of the observed samples that were classified, or predicted, by the algorithm containing the 37 features. The algorithm correctly predicted the difficulty rating for participants on the scale between 2 and 8 with accuracy rates ranging from a low of 22% to a high of 64%. The algorithm appeared to provide better classification accuracy when the observed difficulty rating was 3, 7, or 8 (all above 50%) compared to when the observed difficulty rating was 2, 4, 5, or 6. The response-rating classification accuracy was 0% for predicting a difficulty rating of 1 and 9. It is worth noting that the algorithm was not drastically off the mark; it predicted puzzle rating 2 and 3 instead of 1. Similarly, it predicted 7 or 8 rather than correctly predicting 9.

Emotiv construct data analysis. Raw data for each Emotiv affective construct and each individual difference variable were analyzed to create models with construct features that predicted tasks varying in level of intrinsic load.

- *Preprocessing and feature extraction.* These data were preprocessed and transformed as previously described.
- *Feature selection and decision tree construction.* A Random Forest classification algorithm was applied to the (a) Emotiv construct information, (b) paper-folding scores, and (c) WM capacity scores for 44 participants as a method to classify the self-report rating of difficulty. The self-report ratings of difficulty were aggregated across puzzles, generating a total of 176 responses. There were 11 features extracted from the data that had variable importance values and p -values > 0 . All of the features identified as important were features from Emotiv affective construct data—none were features from the distributions of paper-folding scores or WM capacity scores. Each feature is described in Table 10. The response rating classifications ranged in accuracy from 0% for predicting a difficulty rating of 8 to 83% for predicting a difficulty rating of 9. The algorithm appeared to provide better classification accuracy when the observed difficulty rating was 4, 7, or 9 compared to when the observed difficulty rating was 1, 2, 3, 5, 6, or 8 (see Table 11 for classification accuracy).

Pupil data analysis. An analysis was conducted to determine if any features from the pupil dilation data predicted the self-report difficulty ratings of each puzzle. Unlike the data analysis of the raw EEG and Emotiv constructs, no features were extracted.

Puzzle Performance: Errors and Time

Error and time data were analyzed to answer the fourth research question: Do the number of errors and the amount of time spent on each puzzle increase as the puzzle difficulty increases? As mentioned previously, the number of errors committed and the

amount of time it took to complete each puzzle were logarithmically transformed because their respective distributions were substantially skewed and kurtotic. Two one-way repeated-measures ANOVAs were conducted to evaluate the effect of varying the number of interacting elements across four puzzles on the number of errors committed and the amount of time participants spent completing each puzzle. Results of the analyses for both the average number of errors, Wilks' lambda = .27, $F(3,50) = 44.72$, $p < .001$, multivariate $\eta^2 = .73$, and the average amount of time, Wilks' lambda = .21, $F(3,49) = 62.55$, $p < .001$, multivariate $\eta^2 = .79$, were statistically significant.

Follow-up tests were conducted on the six pairwise comparisons for the average number of errors on each puzzle and for the average amount of time spent on each puzzle. A Holm's sequential Bonferroni procedure was used to control for Type I error with alpha set at .008 (.05/6). Results of each pairwise comparison for the average number of errors committed per puzzle were all statistically significant at $p < .01$, with the exception of the comparison between Puzzle 2 and Puzzle 3, $p = .32$. Likewise, results of each of the pairwise comparisons for the average amount of time spent solving puzzles were all statistically significant at $p < .001$, with the exception of the comparison between Puzzles 2 and 3, $p = .31$. These results suggest, with the exception of Puzzles 2 and 3, that both the average number of errors and the average amount of time spent solving the puzzles increased as the puzzles became more difficult. It appears there was little or no difference between the number of errors on Puzzle 2 and Puzzle 3 and little or no difference between the amount of time spent solving Puzzle 2 and Puzzle 3. The means and standard deviations for errors and time are reported in Tables 12 and 13, respectively. The transformed data are also provided.

Because some violations of parametric test assumptions (e.g., normality) make their conclusions inaccurate, post hoc nonparametric analyses were conducted (Howell, 2002). Nonparametric analyses typically rely on the median, making them less susceptible to outliers that can inflate the variance and bias the mean. Nonparametric analyses were conducted on the median values for the number of errors and for the amount of time spent solving puzzles. Results of this test for both errors, $X^2(3, N = 53) = 53.78, p < .001$, Kendall's $W = .34$, and time, $X^2(3, N = 52) = 76.58, p < .01$, Kendall's $W = .50$, were statistically significant. These nonparametric results replicated the results from the parametric repeated-measures ANOVAs in that they were statistically significant and had large effect sizes.

Follow-up tests were conducted on the six-pairwise comparisons for the median number of errors on each puzzle and for the amount of time spent on each puzzle. A Holm's sequential Bonferroni procedure was used to control for Type I error with alpha set at .008 (.05/6) for each set of tests. Results of each pairwise comparison for the median number of errors were statistically significant, $p < .01$, with the exception of Puzzles 2 and 3, $p = .90$. Similarly, pairwise comparisons of the median amount of time were statistically significant, $p < .001$, with the exception of Puzzles 2 and 3, $p = .53$. These findings indicate that the differences in the median number of errors on each puzzle and the differences in the median amount of time spent solving each puzzle were significantly different and, with the exception of Puzzles 2 and 3, the medians increased as the difficulty—element interactivity—of the puzzles increased. These findings also replicate the follow-up tests conducted on the mean values reported in the parametric

analysis. Table 14 contains the median number of errors and median amount of time spent on each puzzle.

DISCUSSION

Research Questions

Are self-report ratings of cognitive load sensitive to tasks that increase in level of intrinsic load? This effect was statistically significant when the mean CL rating of Puzzle 1, the puzzle with the least element interactivity (the easiest puzzle) was compared to the mean CL ratings of each of the other puzzles, which had higher levels element interactivity. The comparison was also significant when the mean CL rating for Puzzle 2, an easier puzzle, was compared to the mean CL rating for Puzzle 4, the most difficult puzzle, $p < .01$. These findings indicate that participant ratings of CL were sensitive to variations in element interactivity (i.e., intrinsic load), and ratings increased as the difficulty level of the puzzles increased. However, results also suggest that in this environment, self-report CL ratings were less sensitive to the difficulty level of the puzzles when puzzles were somewhat similar in difficulty. For example, when the difference in the element interactivity of two puzzles was smaller—as in Puzzles 2 and 3 or Puzzles 3 and 4—the difference in mean CL ratings was not statistically significant. In general, self-report ratings were sensitive to differentiating the easier puzzles from the more difficult ones but not sensitive for differentiating puzzles that were somewhat similar in difficulty.

Perhaps in this environment self-report CL ratings are sensitive to puzzles differing in element interactivity when the differences are large and reach a certain threshold, as when the two easier puzzles (Puzzles 1 and 2) are compared with the most difficult puzzle (Puzzle 4). It is also possible that the nonsignificant findings in mean CL

ratings for Puzzles 2 and 3 are an indication that there may yet be an undefined aspect of element interactivity—the number of nontarget car moves—that contributes to Puzzle 2 and Puzzle 3 functioning as similarly difficult puzzles.

Are physiological measures sensitive to tasks that increase in level of intrinsic load?

EEG data. The algorithm created with the 29 features extracted from the EEG physiological data appeared to accurately differentiate easier puzzles from the most difficult puzzle. For example, the algorithm accurately classified 55% of Puzzle 1 samples and 52% of Puzzle 2 samples, and it did not classify any of the samples from these two puzzles as Puzzle 4. These findings were consistent with the results of the analysis for self-report CL ratings. However, the EEG data appeared to better distinguish between Puzzles 2 and 3 than did the average self-report CL ratings; over half of Puzzle 2 samples were accurately classified with EEG data, but there was no statistically significant difference in average CL ratings between Puzzles 2 and 3. Neither EEG data nor self-report ratings of CL distinguished well the difficulty levels of Puzzles 3 and 4. Specifically, for Puzzle 4, the algorithm of EEG features accurately classified 45% of the samples but also inaccurately classified 45% of the samples as Puzzle 3. Likewise, the average CL ratings for Puzzles 3 and 4 were statistically indistinguishable. Overall, the results indicate that EEG data collected in this puzzle environment were sensitive to puzzle tasks that increased in level of intrinsic load. Furthermore, with only the exception of Puzzles 3 and 4, the algorithm of extracted EEG features functioned better to distinguish puzzles from one another than did the analyses for self-report CL ratings. It is unclear why the algorithm did not function well to classify Puzzle 3; perhaps participants

experienced the difficulty level of Puzzles 3 and 4 similarly and therefore EEG data were less sensitive to variations of difficulty for these two puzzles.

Emotiv construct data. The algorithm consisting of 14 Emotiv construct features accurately classified approximately 60% to 86% of the samples for Puzzles 1, 2 and 4, indicating the algorithm distinguished well between the difficulty levels of these puzzles. However, the algorithm did not function well to distinguish Puzzle 3 samples from Puzzle 2 and 4 samples, the reason for which is unclear. It is interesting to note that when results from the EEG data analysis and the Emotiv construct data analysis are examined simultaneously, it is evident that the difficulty level of each of the four puzzles is well distinguished. The algorithm of EEG spectral features was most accurate for classifying Puzzle 3, and the algorithm of Emotiv features was most accurate for classifying Puzzles 1, 2, and 4. These results indicate that it is useful to collect and analyze both EEG spectral features and Emotiv construct features to identify algorithms that function well at differentiating tasks varying in difficulty.

Eye tracking data. None of the features from the eye tracking data (pupil dilation, fixation duration, and fixation count) were significant predictors of puzzle difficulty. It was expected that pupil dilation would surface as a significant predictor of puzzle difficulty, as research has shown pupil dilation to increase as task complexity increases (Minassian et al., 2004) and also as the difficulty of a visual search task increases (Porter et al., 2007). There are at least two potential explanations of the lack of findings. Pupil dilation has been reported to be an index of learning, where pupil dilation increases at the start of a learning task but then constricts over the duration of the task as learning occurs (Sibley, Coyne, & Baldwin, 2011). Perhaps the sampling window for which eye-tracking

data were collected and analyzed was too long to detect changes in pupil and eye movement data over each of the puzzle tasks. It may be necessary to use a shorter window of time to detect more subtle changes in pupil dilation, fixation duration, and fixation count.

An alternative explanation is that in the puzzle environment, pupil diameter and eye movement data did not accurately predict puzzle difficulty. Inconsistent findings for the effect of cognitive load on pupil diameter are noted in the literature. For example, Schultheis and Jameson (2004) studied the relationship between pupil diameter and reading task difficulty as a means to assess cognitive load. They measured pupil response while participants read text passages that varied in difficulty (easy and difficult) on a computer screen. The results of the study indicated that although there was a trend toward differences in pupil diameter between the easy and difficulty text passage conditions, it was not statistically significant. The authors hypothesized that pupil size may vary between tasks that are easy and difficult but perhaps only in certain segments of the task (Schulthesis & Jameson, 2004, p. 233). Other researchers have reported findings consistent with this hypothesis. Siegle, Ichikawa, and Steinhauer (2008) studied eye blinks and pupil responses as measures of cognitive load. Their results suggested pupil dilation was indicative of information processing and “sustained cognitive load was accompanied by sustained pupil dilation” (p. 682). Perhaps the segments in the tasks hypothesized by Schultheis and Jameson (2004) where pupil diameter differed between tasks occurred during periods of sustained cognitive load. It is likely that in the puzzle environment the amount of cognitive load imposed by a puzzle task fluctuated for a participant over its duration and capturing such changes with pupil dilation data is

somewhat ineffective. This result suggests that pupil response is not an appropriate measure of task difficulty for this type of puzzle task.

To what extent do objective physiological measures and individual difference variables predict self-report ratings of intrinsic cognitive load? The algorithms for the EEG spectral features and the Emotiv construct features appeared to accurately classify more samples from ratings that were either low (i.e., when students rated the difficulty of the puzzle as a 3 or 4) or high (i.e., when students rated the difficulty of the puzzle as a 7 or 8). These findings are consistent with the results of the self-report CL rating analysis, in that CL ratings were useful for differentiating the easier puzzles (Puzzles 1 and 2), with low ratings, from the most difficult puzzle (Puzzle 4), with high ratings. Pupil-dilation data and eye-movement data did not predict the self-report CL rating of the puzzles. It is possible that pupil response did not predict CL ratings because of the reasons discussed previously.

Individual difference variables (spatial ability and WM capacity) did not predict self-report ratings of intrinsic cognitive load. It appears that in this puzzle environment spatial ability and WMC as measured with paper-folding and an operation span task (administered electronically) did not contribute to ratings of difficulty. These findings may be due to the type of individual difference measures used. For example, the paper-folding task that was used to measure spatial-visualization ability may not be an accurate measure of the spatial-visualization process necessary to solve the tasks used in this puzzle environment. Similarly, there is evidence in the literature indicating that using only a single task of WM capacity is insufficient and the results are “likely to reflect more variance due to specific features of that task than variance due to the construct that

it is meant to measure” (Lewandowsky et al., 2010, p. 577). Using the multiple measures put forth by these researchers may be a promising direction for more accurate measurement of WM capacity.

Do the number of errors and the amount of time spent on each puzzle increase as the puzzle difficulty increases? Results from the analyses indicated that the average number of errors on each puzzle and the average amount of time spent solving each puzzle were statistically different from puzzle to puzzle, except between Puzzles 2 and 3. For the other puzzles, both the average number of errors and the average amount of time spent increased as the difficulty—element interactivity—of the puzzles increased. The insignificant statistical results for Puzzle 2 and Puzzle 3, showing that they did not differ in the average number of errors or the average amount of time spent solving the puzzles, were unexpected. Even though Puzzles 2 and 3 differed in element interactivity—Puzzle 2 required 10 nontarget car moves, and Puzzle 3 required 14 nontarget car moves (see Table 3)—it appears there may be another aspect of element interactivity that could account for the unexpected findings.

Implications

Overall, when studied simultaneously, the algorithms derived from the EEG spectral features and the Emotiv construct features were the most useful of the tested measures for differentiating the difficulty of the puzzles. The algorithm of Emotiv construct features accurately classified approximately 60% or more of the samples for Puzzles 1, 2, and 4 while the algorithm of EEG spectral features accurately classified 71% of the samples for Puzzle 3. It is unclear why Emotiv feature algorithm appears to

accurately classify more puzzle samples than the EEG feature algorithm. Taken together, the algorithms functioned better than self-report ratings to differentiate puzzles.

Integrating digital signals from physiological measures of cognitive load into a learning environment has the potential to create efficient personalized learning experiences by triggering adaptations in the system to respond to the needs of the learner. When sub-optimal states of cognitive load are detected, the system can trigger an adjustment in the learning environment to either reduce the cognitive overload or increase the interest and challenge. This benefit is substantial despite that using data mining methods to derive algorithms of features from big data sets can be time consuming and complex.

While self-report ratings were generally sensitive to the puzzle tasks as they increased in element interactivity. Higher ratings of cognitive load were associated with more difficult puzzles. This result suggests the self-report measure is a reasonably reliable measure of intrinsic cognitive load, which supports this contention by Ayres (2006). However, the self-report measure was not as sensitive for differentiating puzzles that were somewhat similar in difficulty. Unlike the data mined results, self-report CL ratings provided an indication of difficulty, or the amount of element interactivity present in a task. From the CL rating analysis, an inference can be drawn about the difficulty level of a task.

In this puzzle environment, eye tracking data and individual difference variables did not appear to be important variables for differentiating difficulty levels of the puzzles. These results may be due to the sampling techniques and individual difference measures used.

Limitations and Future Directions

The overall goal of this study was to attempt to establish the utility of measuring cognitive load with physiological measures by assessing the responsiveness of self-report and physiological measures in an experimental environment where intrinsic cognitive load (element interactivity) was systematically manipulated. The somewhat inconsistent results from two of the dependent measures underscore the challenge of operationally defining and manipulating element interactivity for a given task. Specifically, the performance data (errors and time) and the average CL self-report ratings for Puzzles 2 and 3 were not statistically different (although CL ratings trended in the expected direction). Even though a small pilot study conducted to test the logic puzzle environment provided preliminary evidence that self-report ratings of CL varied with increasing levels of element interactivity, results with a larger sample indicate perhaps there is another aspect of element interactivity that may have contributed to the unexpected findings. Further exploration of the operational definition of element interactivity (number of nontarget car moves) can shed light on this hypothesis. As in the TOH puzzle, it is likely an aspect of element interactivity involves participants' ability to "inhibit goal compatible but incorrect responses....where the correct subgoal involves moving discs [target and nontarget cars] away from the goal state" (Handley et al., 2002, p. 512). Identifying such recursive moves in the puzzle environment is a first-step toward a deeper understanding of element interactivity. It follows though that individual differences may mediate one's inhibitory processes. As a future direction for this project, it will be beneficial to investigate alternate ways to characterize element interactivity accounting for the effort

required to inhibit “goal compatible but incorrect responses [moves]” (Handley et al., 2002, p. 512) and to control for individual differences in such ability.

Another direction for future research is to replicate the study using other learning stimuli that allows for systematic variation in element interactivity. More specifically, it would be useful to utilize stimuli that is well accepted by the educational psychology research community and that has a distinct definition of element interactivity. Matching tasks (Gevins et al., 1989) and span tasks may be promising stimuli as they typically include a mental manipulation of a specific number of elements.

Finally, conducting a microanalysis of the puzzle data to explore the relationships between specific puzzle strategies and the physiological measures may help identify the segments in the puzzles that are perceived as easy and difficult, and relate those experiences to the physiological measures. As mentioned, even though each puzzle can be characterized on a scale of difficulty from easy to difficult, it is likely that each participant experiences variations of difficulty within each puzzle. The physiologic measures provide continuous time series data that lend themselves well to this type of microanalysis. These types of future studies will help identify and link participants’ experience of cognitive load with physiological indicators and thus facilitate the opportunity to use such digital input to manipulate learning environments according the level of cognitive load experienced. However, a post-task self-report measure of CL appears to have reliably functioned in the puzzle environment to detect variations in element interactivity when the variation is substantial. It is beneficial for the research community to have access to a variety of methods for measuring cognitive load and an understanding of best-practices for each method.

REFERENCES

- Amadiou, F., van Gog, T., Paas F., Tricot, A., & Mariné, C. (2009). Effects of prior knowledge and concept-map structure on disorientation, cognitive load, and learning. *Learning and Instruction, 19*, 376–386.
- Anderson, E. W., Potter, K. C., Matzen, L. E., Shepherd, J. F., Preston, G. A., & Silva, C. T. (2011). A user study of visualization effectiveness using EEG and cognitive load. *Computer Graphics Forum, 30*, 791–800. doi: 10.1111/j.1467-8659.2011.01928.x
- Andreassi, J. L. (2000). *Psychophysiology: human behavior and physiological response*. Mahwah, NJ; London: L. Erlbaum.
- Antonenko, P., Paas, F., Grabner, R., & van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review, 22*, 425–438.
- Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction, 16*, 389–400.
- Ayres, P., & van Gog, T. (2009). State of the art research into cognitive load theory. *Computers in Human Behavior, 25*, 253–257.
- Baddeley, A. (2007). *Working memory, thought, and action*. New York, NY: Oxford University Press.
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191–215.
- Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. C. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307-337). Greenwich, CT: Information Age.
- Baydogan, M., Runger, G., & Atkinson, R. K. (in press). *Knowledge discovery with sensor data*.
- Baydogan, M., Runger, G., & Tuv, E. (2011, November). *A bag-of-features framework to classify time series*. Paper presented at Institute for Operations Research and the Management Sciences, Charlotte, North Carolina.
- Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276–292.
- Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., . . . Craven, P. (2007). EEG correlates of task engagement and mental workload in

- vigilance, learning and memory tasks. *Aviation, Space and Environmental Medicine*, 78(5), B231–B244.
- Bobrov, P., Frolov, A., Cantor, C., Fedulova, I., Bakhnyan, M., & Zhavoronkov, A. (2011). Brain-computer interface based on generation of visual images. *PLoS ONE* 6(6): e20674. doi:10.1371/journal.pone.0020674
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10, 433–436.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Brünken, R., Steinbacher, S., Plass, J., & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology*, 49(2), 109–119.
- Brünken, R., Plass, J., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 53–61.
- Brünken, R., Plass, J., & Leutner, D. (2004). Assessment of cognitive load in multimedia learning with dual-task methodology: Auditory load and modality effects. *Instructional Science*, 32, 115–132.
- Campbell, A., Choudhury, T., Hu, S., Hong, L., Mukerjee, M., Rabbi, M., & Raizada, R. (2010). NeuroPhone: Brain-mobile phone interface using a wireless EEG headset. In *Proceedings of the Second ACM SIGCOMM Workshop on Networking, Systems, and Applications on Mobile Handhelds* (pp. 3–8). New York, NY: ACM
- Carroll, J. B. (1974). *Psychometric tests as cognitive tasks: A new “structure of intellect.”* (Technical Report No. 4). Princeton, NJ: Educational Testing Service.
- Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior*, 25, 315–324.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression / correlation analysis for behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum.
- Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user’s guide. *Psychonomic Bulletin & Review*, 12(5), 769–786.
- de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38, 105–134.
- DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, 100(1), 223–234.

- DeStefano, D., & LeFevre, J. A. (2007). Cognitive load in hypertext reading: A review. *Computers in Human Behavior, 23*, 1616–1641.
- Doolittle, P. E., & Mariano, G. M. (2008). Working memory capacity and mobile multimedia learning environments: Individual differences in learning while mobile. *Journal of Educational Multimedia and Hypermedia, 17*(4), 511–530.
- Duchowski, A. T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, 34*, 455–470.
- Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). *Manual for kit of factor referenced cognitive tests*. Princeton, NJ: Educational Testing Service.
- Emotiv [Apparatus and software]. (2010). San Francisco, CA: Emotive Systems.
- Engle, R. W., Cantor, J., & Carullo, J. J. (1992). Individual differences in working memory and comprehension: A test of four hypotheses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(5), 972–992.
- Gerjets, P., Scheiter, K., & Cierniak, G. (2009). The scientific value of cognitive load theory: A research agenda based on the structuralist view of theories. *Educational Psychology Review, 21*, 43–54.
- Gerjets, P., Scheiter, K., Opfermann, M., Hesse, F. W., & Eysink, T. H. S. (2009). Learning with hypermedia: The influence of representational formats and different levels of learner control on performance and learning behavior. *Computers in Human Behavior, 25*, 360–370.
- Gevins, A., & Smith, M. E. (2003). Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science, 4*(1-2), 113-131.
- Gevins, A., Smith, M. E., Leong, H., McEvoy, L., Whitfield, S., Du, R., & Rush, G. (1989). Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. *Human Factors, 40*(1), 79–91.
- Gonzalez-Sanchez, J., & Chavez-Echeagaray, M. E. (2011). Timestamp synchronization and file compiler (Version 2) [Computer software]. Tempe: Arizona State University.
- Gonzalez-Sanchez, J., Chavez-Echeagaray, M. E., Atkinson, R., & Burleson, B. (2011, June). *ABE: An agent-based software architecture for a multimodal emotion recognition framework*. Paper presented at the Ninth Working IEEE/IFIP Conference on Software Architecture, Boulder, CO. Abstract retrieved from <http://doi.ieeecomputersociety.org/10.1109/WICSA.2011.32>

- Gyselinck, V., Jamet, E., & DuBois, V. (2008). The role of working memory components in multimedia comprehension. *Applied Cognitive Psychology, 22*, 353–374.
- Handley, S. J., Capon, A., Copp, C., & Harper, C. (2002). Conditional reasoning and the Tower of Hanoi: The role of spatial and verbal working memory. *British Journal of Psychology, 93*(4), 501–518.
- Hearn, R. (2006). *Games, puzzles and computation* (Doctoral thesis). Retrieved from <http://dspace.mit.edu/handle/1721.1/37913>
- Hearn, R. (2009). Subway Shuffle [Computer software]. Hanover, NH: Author.
- Höffler, T. N. (2010). Spatial ability: Its influence on learning with visualizations: A meta-analytic review. *Educational Psychology Review, 22*, 245–269.
- Höffler, T. N., Sumfleth, E., & Leutner, D. (2006). The role of spatial ability when learning from an instructional animation or a series of static pictures. In J. Plass (Ed.), *Proceedings of the NYU Symposium on Technology and Learning*. New York: New York University. Retrieved from http://create.alt.ed.nyu.edu/symposium2006/NYUSymposium2006_Hoeffler_Sumfleth_Leutner.pdf
- Hordern, E. (1986). *Sliding piece puzzles*. New York, NY: Oxford University Press.
- Howell, D. C. (2009). *Statistical methods for psychology* (7th ed.). Pacific Groove, CA: Wadsworth.
- Kalyuga, S. (2005). Prior knowledge principle in multimedia learning. In R. Mayer (ed.), *Cambridge handbook of multimedia learning* (pp. 325–337). New York, NY: Cambridge University Press.
- Kalyuga, S. (2007). Enhancing instructional efficiency of interactive e-learning environments: A cognitive load perspective. *Educational Psychology Review, 19*, 387–399.
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1), 23–31.
- Kalyuga, S., Chandler, P., & Sweller, J. (1998). Levels of expertise and instructional design. *Human Factors, 40*, 1–17.
- Klimesch, W., Schack, B., & Sauseng, P. (2006). The functional significance of theta and upper alpha oscillations. *Experimental Psychology, 52*(2), 99–108.
- Klimesch, W., Schimke, H., & Pfurtscheller, G. (1993). Alpha frequency, cognitive load and memory performance. *Brain Topography, 5*(3), 241–251.

- Klingner, J. (2010). *Measuring cognitive load during visual tasks by combining pupillometry and eye tracking*. (Doctoral dissertation, Stanford University). Retrieved from <http://graphics.stanford.edu/~klingner/publications/index.html>
- Lee, H., Plass, J. L., & Homer, B. D. (2006). Optimizing cognitive load for learning from computer-based science simulations. *Journal of Educational Psychology, 98*(4), 902–913.
- Lewandowsky, S., Oberauer, K., Yang, L., & Ecker, U. K. H. (2010). *A working memory test battery for MATLAB, 42*(2), 571–585.
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychology, 38*(1), 43–52.
- Minassian, A., Granholm, E., Verney, S., & Perry, W. (2004). Pupillary dilation to simple vs. complex tasks and its relationship to thought disturbance in schizophrenia patients. *International Journal of Psychophysiology, 52*, 53–62.
- Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Burlington, MA: Elsevier.
- Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429–434.
- Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1–4.
- Paas, F., Tuovinen, J. E., Tabbers, H., & van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63–71.
- Paas, F., van Gog, T., & Sweller, J. (2010). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review, 22*, 115–121.
- Paas, F., & van Merriënboer, J. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*, 122–133.
- Paas, F., van Merriënboer, J., & Adam, J. (1994). Measurement of cognitive load in instructional research. *Perception and Motor Skills, 79*, 419–430.
- Pajares, F. (2002). *Overview of social cognitive theory and self-efficacy*. Unpublished manuscript, Emory University, Atlanta, GA. Retrieved from <http://www.emory.edu/EDUCATION/mfp/eff.html>

- Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology, 86*(2), 193–203.
- Pelli, D. G. (1997). The Video Toolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.
- Pivik, R. T., Broughton, R. J., Coppola, R., Davidson, R. J., Fox, N., & Nuwer, M. R. (1993). Guidelines for the recording and quantitative analysis of electroencephalographic activity in research contexts. *Psychophysiology, 30*, 547-558.
- Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and counting: Insights from pupillometry. *The Quarterly Journal Of Experimental Psychology, 60*(2), 211–229.
- QuicKeys [Computer software]. West Des Moines, IA: StartleyTechnologies.
- Schink Joseph, S., & Atkinson, R. K. (2010, May). *Self-reported intrinsic cognitive load and variations in element interactivity*. Poster presented at 2010 American Educational Research Association Annual Meeting, Denver, CO.
- Schmeichel, B. J., Volokhov, R. N., & Demaree, H. A. (2008). Working memory capacity and the self-regulation of emotional expression and experience. *Journal of Personality and Social Psychology, (95)*6, 1526–1540.
- Schnotz, W., & Kirschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychological Review, 19*, 469-508.
- Schultheis, H., & Jameson, A. (2004). Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioral methods. In W. Nejdl & P. De Bra (Eds.), *Adaptive hypermedia and adaptive web-based systems: Proceedings of AH 2004* (pp. 225–234). Berlin, Germany: Springer-Verlag.
- Sibley, C., Coyne, J., Baldwin, C. (2011). Pupil dilation as an index of learning. *Proceedings of 55th Annual Meeting of the Human Factors and Ergonomics Society, 237-241*.
- Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before you think: Blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology, 45*, 679–687.
- Smith, M. E., Gevins, A., Brown, H., Karnik, A., & Du, R. (2001). Monitoring task loading with multivariate EEG measures during complex forms of human-computer interaction. *Human Factors, 43*(3), 366–380.

- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12, 257–285.
- Sweller, J. (2005). Implications for cognitive load in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 19–30). New York, NY: Cambridge University Press.
- Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, 22, 123–138.
- Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston MA: Pearson Education.
- Tobii Studio (2009) [Computer software]. Stockholm, Sweden: Tobii Technology.
- Tobii Technology (2009) [Apparatus]. Stockholm, Sweden: Tobii Technology.
- Tsianos, N., Germanakos, P., Lekkas, Z., Mourlas, C., & Samaras, G. (2010). Working memory span and e-learning: The effect of personalization techniques on learners' performance. In P. De Bra, A. Kobsa, & D. Chin (Eds.), *Lecture notes in computer science, vol. 6075: User modeling, adaptation, and personalization* (pp. 64–74). Berlin, Germany: Springer-Verlag.
- Tuv, E., Borisov, A, Runger, G., & Torkkola, K. (2009). Feature selection with ensembles, artificial variables, and redundancy elimination. *Journal of Machine Learning Research*, 10, 1239–1263.
- Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37, 498-505.
- van Gerven, P. W. M., Paas, F., van Merriënboer, J. J. G., & Schmidt, H. G. (2002). Cognitive load theory and aging: Effects of worked examples on training efficiency. *Learning and Instruction*, 12, 87–105.
- van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, 17(2), 147–177.
- Verney, S. P., Granholm, E., & Dionisio, D. P. (2001). Pupillary responses and processing resources on the visual backward task. *Psychophysiology*, 38, 76–83.
- Whelan, R. (2007). Neuroimaging of cognitive load in instructional media. *Educational Research Review*, 2, 1–12.
- Wouters, P., Paas, F., & van Merriënboer, J. J. G. (2008). How to optimize learning from animated models: A review of guidelines based on cognitive load. *Review of Educational Research*, 78(3), 645–675.

Table 1

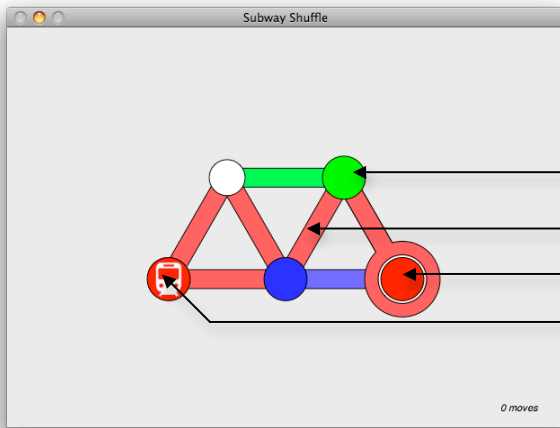
Description of Dependent/Process Measures used to Assess Cognitive Load

| Description | Source |
|--|--------------------|
| Self-report | |
| Perceived difficulty | Self-report survey |
| Puzzle performance | |
| Time | Log file |
| Errors (number of moves over minimum) | Log file |
| Physiological | |
| Fixation duration | Eye tracker |
| Fixation count | Eye tracker |
| Pupil dilation (left/right) | Eye tracker |
| EEG node data (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4) | EPOC headset |
| EPOC affective construct data (excitement, engagement, boredom, frustration, and meditation) | EPOC headset |
| Head movement (GYROX, GYROY) | EPOC headset |

Table 2

Puzzle Description

| Puzzle | Level | Puzzle elements | | | |
|--------|-------|-----------------|--------|----------|--------|
| | | Cars | Tracks | Stations | Colors |
| 1 | 1 | 5 | 9 | 6 | 4 |
| 2 | 2 | 5 | 9 | 6 | 4 |
| 3 | 3 | 5 | 6 | 6 | 3 |
| 4 | 4 | 6 | 9 | 7 | 3 |



- moveable colored token or subway car (nontarget car)
- subway track
- destination station with red ring
- red subway car (target car)

Figure 1. Image of practice puzzle C with labeled components. This puzzle contains four target cars (one target car and three nontarget cars), seven tracks, five stations, and three colors.

Table 3

Puzzle Difficulty: Target Car and Nontarget Car Moves

| Puzzle | Level | Puzzle Moves | | |
|--------|-------|--------------------|------------|---------------|
| | | Moves ^a | Target car | Nontarget car |
| 1 | 1 | 11 | 3 | 8 |
| 2 | 2 | 15 | 5 | 10 |
| 3 | 3 | 19 | 5 | 14 |
| 4 | 4 | 23 | 5 | 18 |

^a The minimum number of moves required to solve the puzzle.

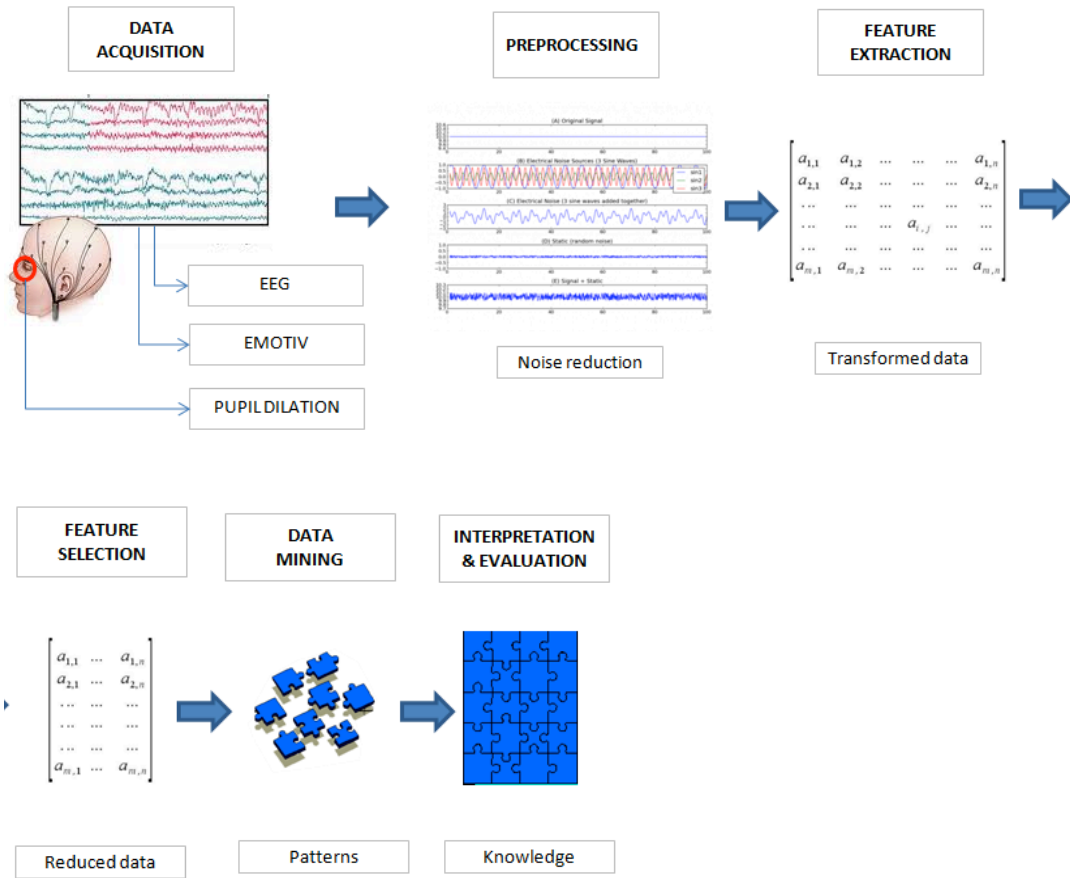


Figure 2. The data mining process used to discover the sensor features important for predicting the difficulty level of the puzzle and for predicting the self-report difficulty ratings of each puzzle (reprinted with permission; Baydogan, Runger, & Atkinson, in press).

Table 4

EEG Spectral Features

| Features ^a | Min p value | Final importance |
|-----------------------|---------------|------------------|
| T8_theta_min | 0 | 100% |
| O1_theta_min | 0 | 96.23% |
| O2_theta_min | 0 | 90.99% |
| AF3_theta_min | 0 | 85.76% |
| F4_alpha_min | 0.00452485 | 84.62% |
| F4_theta_min | 3.82E-07 | 78.29% |
| O1_alpha_min | 0 | 77.27% |
| F8_theta_min | 8.69E-07 | 75.10% |
| AF3_alpha_min | 4.24E-06 | 74.69% |
| F4_beta_min | 1.35E-07 | 71.94% |
| FC6_beta_min | 0 | 71.46% |
| O2_alpha_min | 0.018185 | 69.14% |
| P8_theta_max | 7.67E-07 | 67.67% |
| O2_alpha_max | 0.0433842 | 65.38% |
| O2_beta_min | 2.50E-05 | 65.20% |
| FC6_theta_min | 0.000467568 | 63.20% |
| F8_alpha_min | 2.65E-07 | 62.05% |
| FC5_theta_min | 0.00523358 | 61.82% |
| T7_alpha_min | 3.82E-06 | 61.45% |
| P8_alpha_max | 0.000274333 | 58.41% |
| AF4_theta_max | 2.31E-05 | 57.97% |
| T7_theta_min | 0.00155745 | 57.83% |
| P7_theta_min | 0 | 57.03% |
| F7_theta_min | 0.0236898 | 54.84% |
| T8_alpha_max | 0.00198814 | 53.34% |
| FC5_alpha_min | 0.00022344 | 50.77% |
| O2_theta_max | 0.0034853 | 50.25% |

| | | |
|---------------|-----------|--------|
| F7_beta_min | 0.0355958 | 48.56% |
| FC6_theta_max | 0.0167877 | 47.62% |

^a The features are identified as *electrode placement_wave type_min or max*. For example, “F7_theta_min” identifies the minimum value of theta activity associated with the frontal F7 node and “AF4_theta_max” identifies the maximum theta activity at the anterior frontal AF4 node.

Table 5

Confusion Matrix for EEG Spectral Features

| Observed puzzle | Predicted puzzle | | | |
|-----------------|------------------|-----|-----|-----|
| | 1 | 2 | 3 | 4 |
| 1 | 55% | 39% | 7% | - |
| 2 | 7% | 52% | 41% | - |
| 3 | 2% | 18% | 71% | 9% |
| 4 | - | 9% | 45% | 45% |

Table 6

Emotiv Features and p Values for Predicting Puzzle

| Feature | <i>p</i> value |
|--------------------------------|----------------|
| Frustration_kurtosis | 0.00 |
| Engagement/Boredom_mean | 0.000 |
| Engagement/Boredom_max | 0.000 |
| Engagement/Boredom_median | 0.037 |
| Engagement/Boredom_min | 0.000 |
| Short Term Excitement_kurtosis | 0.000 |
| Short Term Excitement_max | 0.026 |
| Long Term Excitement_variance | 0.000 |
| Short Term Excitement_min | 0.000 |
| Long Term Excitement_min | 0.000 |
| Short Term Excitement_skew | 0.000 |
| Long Term Excitement_skew | 0.000 |
| Meditation_kurtosis | 0.000 |
| Frustration_variance | 0.000 |

Table 7

Confusion Matrix: Emotiv Features for Predicting Puzzle

| Observed puzzle | Predicted puzzle ^a | | | |
|-----------------|-------------------------------|-----|-----|-----|
| | 1 | 2 | 3 | 4 |
| 1 | 59% | 27% | 2% | 11% |
| 2 | - | 86% | 2% | 12% |
| 3 | 5% | 36% | 23% | 36% |
| 4 | - | 20% | 5% | 75% |

^a $n = 44$.

Table 8

EEG Spectrum Features for Predicting Difficulty Ratings

| Features ^a | Min p value | Final importance |
|-----------------------|---------------|------------------|
| FC6_beta_min | 0 | 100% |
| T7_beta_min | 0 | 86.51% |
| F7_beta_min | 0 | 80.75% |
| F3_beta_min | 0 | 79.63% |
| AF3_theta_min | 0 | 77.81% |
| F7_alpha_min | 0 | 74.84% |
| T8_beta_min | 0 | 72.82% |
| FC5_theta_min | 0 | 70.81% |
| F8_beta_min | 9.03E-08 | 68.00% |
| AF4_theta_min | 0 | 67.61% |
| FC5_beta_min | 7.01E-05 | 66.65% |
| F7_beta_mean | 0.000119819 | 63.49% |
| O2_theta_min | 0 | 56.36% |
| P7_beta_min | 4.08E-05 | 54.98% |
| O2_beta_min | 2.49E-07 | 54.44% |
| FC5_theta_max | 0 | 54.02% |
| T7_beta_mean | 0.00537507 | 53.74% |
| F8_theta_min | 3.52E-06 | 52.10% |
| FC6_alpha_max | 3.21E-07 | 51.59% |
| F4_theta_min | 0 | 50.28% |
| F3_theta_max | 2.11E-08 | 49.95% |
| F3_theta_min | 6.36E-05 | 49.31% |
| AF3_alpha_min | 9.91E-05 | 49.17% |
| F4_beta_min | 1.66E-06 | 48.06% |
| T8_theta_min | 0.000471057 | 47.90% |
| F4_theta_max | 0 | 47.90% |
| F3_alpha_max | 1.99E-07 | 47.83% |

| | | |
|---------------|-------------|--------|
| FC6_theta_min | 3.50E-08 | 46.95% |
| AF4_alpha_max | 0.0232589 | 45.97% |
| O2_alpha_min | 1.46E-06 | 45.04% |
| O1_beta_min | 3.75E-09 | 44.23% |
| P7_alpha_max | 0 | 44.01% |
| O1_theta_min | 0.000968231 | 40.01% |
| FC6_alpha_min | 8.18E-05 | 39.34% |
| P7_theta_max | 0.0299127 | 37.05% |
| T7_theta_max | 0.0166537 | 35.28% |
| O1_theta_max | 0.0363198 | 30.55% |

^a The features are identified as *electrode placement_wave type_min or max*. For example, “F7_theta_min” identifies the minimum value of theta activity associated with the frontal F7 node and “AF4_theta_max” identifies the maximum theta activity at the anterior frontal AF4 node.

Table 9

Confusion Matrix: EEG Spectral Features for Predicting Self-report Difficulty Ratings

| Observed difficulty rating | Predicted difficulty rating ^a | | | | | | | | |
|----------------------------|--|-----|-----|-----|-----|-----|-----|-----|----|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 0% | 79% | 21% | | | | | | |
| 2 | | 35% | 53% | 12% | | | | | |
| 3 | | 4% | 57% | 35% | 4% | | | | |
| 4 | | 3% | 17% | 45% | 24% | 7% | 3% | | |
| 5 | | | 17% | 43% | 22% | 17% | | | |
| 6 | | | 4% | 22% | 17% | 22% | 35% | | |
| 7 | | | | 3% | 3% | 20% | 63% | 10% | |
| 8 | | | | | | | 36% | 64% | |
| 9 | | | | | | | 33% | 67% | 0% |

Note. The predictions were from 44 participants aggregated across the four puzzles. Participants rated their perception of puzzle difficulty on a scale from 1 (*not at all difficult*) to 9 (*extremely difficult*).

^a $n = 176$.

Table 10

Emotiv Features and p Values for Predicting Self-report Rating of CL

| Feature ^a | <i>p</i> value |
|--------------------------------|----------------|
| Short Term Excitement_max | 0.000 |
| Short Term Excitement_variance | 0.000 |
| Short Term Excitement_mean | 0.000 |
| Engagement/Boredom_max | 0.000 |
| Engagement/Boredom_variance | 0.000 |
| Long Term Excitement_max | 0.000 |
| Short Term Excitement_median | 0.000 |
| Long Term Excitement_variance | 0.000 |
| Engagement/Boredom_min | 0.000 |
| Frustration_median | 0.000 |
| Frustration_max | 0.000 |

^a The features are identified as *Emotiv construct_min or max*. For example, “Short Term Excitement_max” identifies the maximum value of Short Term Excitement associated with the construct.

Table 11

Confusion Matrix: Emotiv Features for Predicting Self-report Rating of CL

| Observed difficulty rating | Predicted difficulty rating ^a | | | | | | | | |
|----------------------------|--|----|-----|-----|-----|----|-----|----|-----|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | 29% | 2% | 14% | 36% | | | | | |
| 2 | 18% | 6% | 24% | 47% | 6% | | | | |
| 3 | | 4% | 22% | 61% | 13% | | | | |
| 4 | | 7% | 7% | 64% | 18% | 4% | | | |
| 5 | | | 4% | 48% | 26% | 4% | 17% | | |
| 6 | | | | 26% | 43% | 4% | 26% | | |
| 7 | | | | 3% | 20% | 3% | 70% | | 3% |
| 8 | | | | | | 9% | 82% | 0% | 9% |
| 9 | | | | | | | 17% | | 83% |

Note. The predictions were from 44 participants aggregated across the four puzzles. Participants rated their perception of puzzle difficulty on a scale from 1 (*not at all difficult*) to 9 (*extremely difficult*).

^a $n = 176$.

Table 12

Puzzle Errors: Means and Standard Deviations for Original and Transformed Data

| Puzzle ^a | Original data | | Transformed data | |
|---------------------|---------------|-------|-------------------|------|
| | Mean | SD | Mean ^b | SD |
| 1 | 2.79 | 3.79 | .41 | 0.38 |
| 2 | 14.06 | 16.08 | .93 | 0.51 |
| 3 | 13.58 | 16.80 | .82 | 0.61 |
| 4 | 28.19 | 31.51 | 1.22 | 0.50 |

^a $n = 53$.^b Logarithmic units.

Table 13

Time to Solve Puzzle: Means and Standard Deviations for Original and Transformed Data

| Puzzle ^a | Original data | | Transformed data | |
|---------------------|-------------------|-------|-------------------|------|
| | Mean ^b | SD | Mean ^c | SD |
| 1 | 41.39 | 23.12 | 1.55 | 0.30 |
| 2 | 93.28 | 63.76 | 1.87 | 0.36 |
| 3 | 88.97 | 73.35 | 1.80 | 0.46 |
| 4 | 145.12 | 87.29 | 2.07 | 0.37 |

^a $n = 52$.

^b Seconds.

^c Logarithmic units.

Table 14

Median Values for Puzzle Errors and Time

| Puzzle | Errors ^a | Time ^b |
|--------|---------------------|-------------------|
| 1 | 2 | 34.88 |
| 2 | 10 | 74.28 |
| 3 | 6 | 67.15 |
| 4 | 15 | 119.26 |

^a $n = 53$.

^b $n = 52$.

Table 15

Self-report Ratings of Cognitive Load

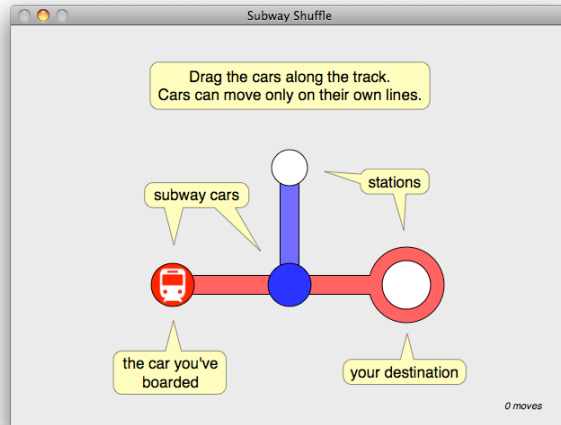
| Puzzle ^a | Mean | SD |
|---------------------|------|-------|
| 1 | 3.31 | 1.725 |
| 2 | 5.19 | 2.047 |
| 3 | 5.28 | 2.149 |
| 4 | 6.02 | 2.042 |

Note. Cognitive load measure is self-reported rating on difficulty scale from 1 (*not at all difficult*) to 9 (*extremely difficult*).

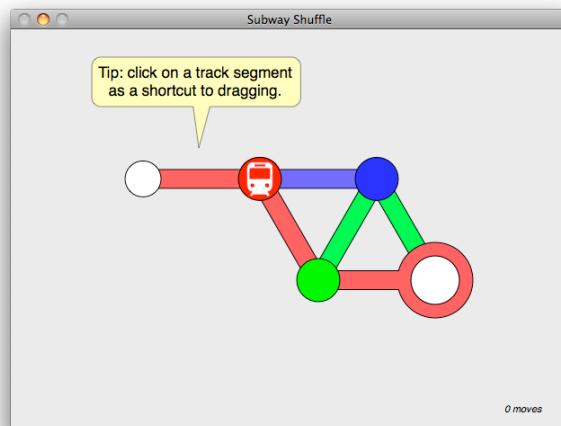
^a *n* = 54.

APPENDIX A
PRACTICE PUZZLES

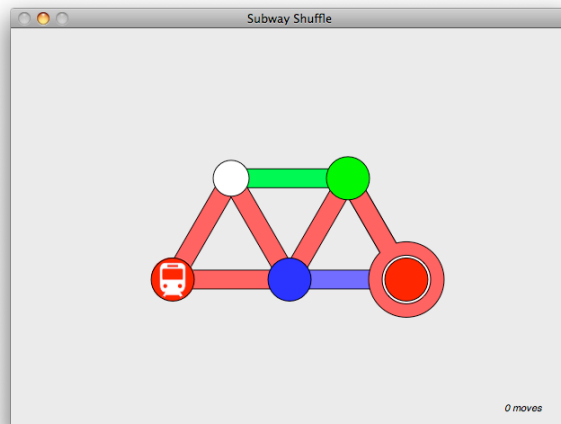
Puzzle A - 3 moves



Puzzle B - 10 moves

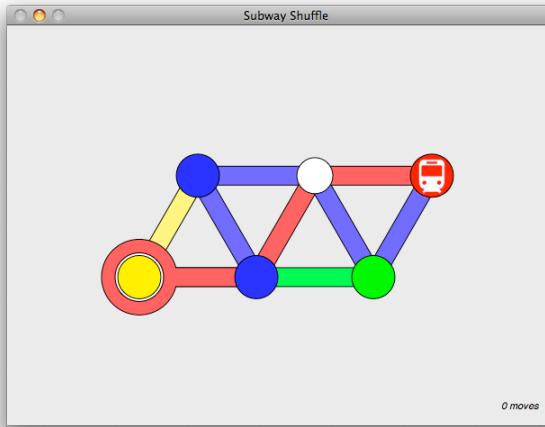


Puzzle C - 12 moves

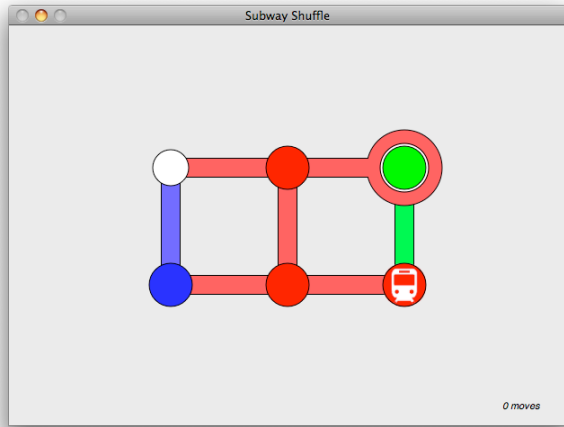


APPENDIX B
EXPERIMENTAL PUZZLES

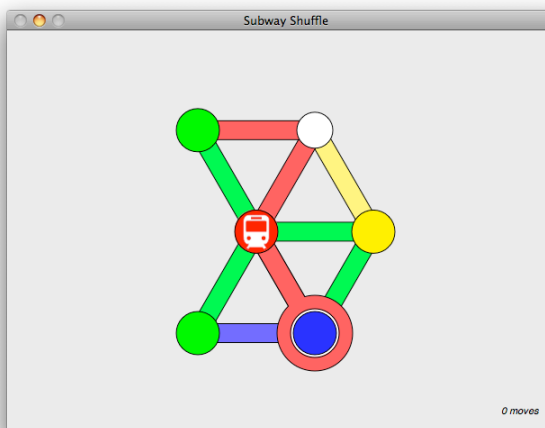
Puzzle 1: level 1 - 11 moves



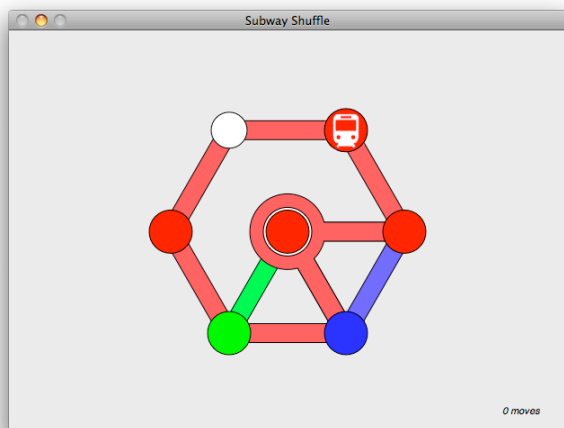
Puzzle 3: level 3 - 19 moves



Puzzle 2: level 2 - 15 moves



Puzzle 4: level 4 - 23 moves



APPENDIX C
PARTICIPANT DATA SURVEY

1. Subject ID:

*

2. Age:

3. Gender:

- Male
 Female

4. Race/ Ethnicity:

- Asian/Pacific Islander Black/African-American Caucasian Hispanic Native American/Alaska Native Other/Multi-Racial Decline to Respond
-

5. What is your primary language?

- English
 Spanish
 Other

6. Vision

| | Yes | No |
|---|-----------------------|-----------------------|
| Do need glasses or contacts to correct your vision? * | <input type="radio"/> | <input type="radio"/> |
| Are you wearing contacts or glasses now? * | <input type="radio"/> | <input type="radio"/> |
| Do you have any degree of color-blindness? * | <input type="radio"/> | <input type="radio"/> |

7. Please indicate which non-verbal logic puzzles you've played in the past 6 months. These types of puzzles require you to manipulate shapes or patterns to reach the solution. The format of the puzzles may be paper-based,

electronic or have hand-held puzzle pieces. *

| | Yes | No |
|--|-----------------------|-----------------------|
| Tetris | <input type="radio"/> | <input type="radio"/> |
| Sudoku | <input type="radio"/> | <input type="radio"/> |
| Puzzles with movable pieces (Rubik's cube, Traffic jam, Tipover, Blokus) | <input type="radio"/> | <input type="radio"/> |
| Jigsaw puzzles | <input type="radio"/> | <input type="radio"/> |
| Chess | <input type="radio"/> | <input type="radio"/> |
| Tangrams | <input type="radio"/> | <input type="radio"/> |

Please list any other non-verbal logic puzzles that you've played in the past 6 months:

1

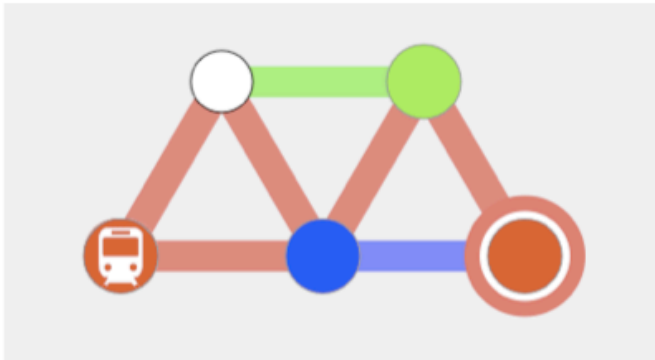
2

3

How often have you played these kinds of puzzles in the past 6 months?

never a few times a few times per month a few times per week

8. Have you played a puzzle game that looks like this?



*

Yes
 No

How often have you played?

never a few times per year a few times per month a few times per week daily

APPENDIX D

COGNITIVE LOAD & PUZZLE SELF-EFFICACY SURVEY

1. How difficult was this puzzle to solve?

| | | | | | | | | | | |
|-------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|------------------------|
| not at all difficult | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | extremely difficult |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |

2. How confident are you that you can solve more puzzles like these?

| | | | | | | | | | | |
|-------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|------------------------|
| not at all confident | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | extremely confident |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |

3. How interested were you in finishing this puzzle?

| | | | | | | | | | | |
|--------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------|
| not at all interested | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | extremely interested |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |

4. How stressed were you while working on this puzzle?

| | | | | | | | | | | |
|------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| not at all stressed | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | extremely stressed |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |

5. How certain are you that you can solve more puzzles like these?

| | | | | | | | | | | |
|------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|----------------------------|
| can not do it | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | highly certain I can |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |

6. How relaxed were you while working on this puzzle?

| | | | | | | | | | | |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|----------------------|
| not at all relaxed | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | extremely relaxed |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |

7. How frustrated were you while working on this puzzle?

| | | | | | | | | | | |
|--------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------|
| not at all frustrated | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | extremely frustrated |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |

APPENDIX E

EXIT SURVEY

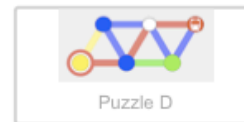
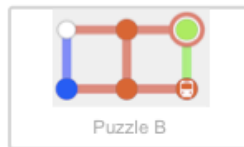
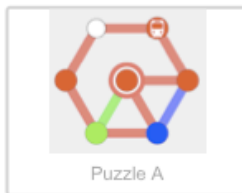
1. Subject ID:

*

2. Rank the puzzles you worked on from easy to hard (see puzzle images below for reference).

Drag items from the left-hand list into the right-hand list to order them.

| | |
|---|--|
| A | |
| B | |
| C | |
| D | |



3. Overall, rate the difficulty level of the puzzles.

| | | | | | | | | | | |
|-----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|----------------|
| very easy | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | very difficult |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |

4. If you were to do another puzzle activity similar to this, would you prefer to solve puzzles that were:

| | | | | |
|-----------------------|-----------------------|-----------------------|-------------------------|-----------------------|
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| a lot easier | a little easier | exactly the same | a little more difficult | a lot more difficult |

5. Overall, the puzzles were:

| | | | | | | | | | |
|--------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------|
| way too hard | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | way too easy |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

6. The activity would have been more enjoyable if the puzzles were:



| | | | | |
|-----------------------|-------------------------|-----------------------|-----------------------|-----------------------|
| <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| much more difficult | a little more difficult | the same | a little easier | much easier |

APPENDIX F
IRB APPROVAL



Office of Research Integrity and Assurance

To: Robert Atkinson
BY M1-08

From:  Mark Roosa, Chair 
Soc Beh IRB

Date: 02/21/2012

Committee Action: **Renewal**

Renewal Date: 02/21/2012

Review Type: Expedited F4 F7

IRB Protocol #: 1102006041

Study Title: Measuring Cognitive Load

Expiration Date: 02/16/2013

The above-referenced protocol was given renewed approval following Expedited Review by the Institutional Review Board.

It is the Principal Investigator's responsibility to obtain review and continued approval of ongoing research before the expiration noted above. Please allow sufficient time for reapproval. Research activity of any sort may not continue beyond the expiration date without committee approval. Failure to receive approval for continuation before the expiration date will result in the automatic suspension of the approval of this protocol on the expiration date. Information collected following suspension is unapproved research and cannot be reported or published as research data. If you do not wish continued approval, please notify the Committee of the study termination.

This approval by the Soc Beh IRB does not replace or supersede any departmental or oversight committee review that may be required by institutional policy.

Adverse Reactions: If any untoward incidents or severe reactions should develop as a result of this study, you are required to notify the Soc Beh IRB immediately. If necessary a member of the IRB will be assigned to look into the matter. If the problem is serious, approval may be withdrawn pending IRB review.

Amendments: If you wish to change any aspect of this study, such as the procedures, the consent forms, or the investigators, please communicate your requested changes to the Soc Beh IRB. The new procedure is not to be initiated until the IRB approval has been given.

Please retain a copy of this letter with your approved protocol.

APPENDIX G
LICENSING AGREEMENT

LICENSING AGREEMENT

THIS AGREEMENT, entered into as of March 30, 2011 between Educational Testing Service (hereinafter called "ETS"), a nonstock, nonprofit corporation organized and existing under the Education Law of the State of New York, with offices at Princeton, New Jersey 08541, and

Stacey Schink Joseph
1000 S Forest Mall
Payne Building #127
Tempe, AZ 85287

(Hereinafter called "Licensee"),

WITNESSETH:

WHEREAS, ETS is the publisher and copyright owner of certain test materials;
and

WHEREAS, Licensee wishes to produce editions of:

VZ-2 Paper Folding Test

NOW, THEREFORE, ETS agrees that Licensee may reproduce and distribute 65 copies of each of the above edition for use in a research study, subject to the following terms and conditions:

1. Each copy of any edition produced under the Agreement shall bear a copyright notice exactly as it appears on the original test, followed by the statement:
Reproduced under license.
2. Licensee agrees to pay ETS a licensing fee of \$40.00. Payment shall be sent to Educational Testing Service to the attention of:

Kristina Phillips M/S 42-L
Assistant, Copyright Group
Office of General Counsel
Rosedale Road
Princeton, New Jersey 08541

Licensee will be responsible for any costs involved in the composition, reproduction, and distribution of the editions licensed herein.



